**In the Name of God**


# Journal of
## Information Systems & Telecommunication
**Vol. 6, No. 2, April-June 2018, Serial Number 22**

**Research Institute for Information and Communication Technology**
**Iranian Association of Information and Communication Technology**
**Affiliated to: Academic Center for Education, Culture and Research (ACECR)**




**Print ISSN**: 2322-1437
**Online ISSN**: 2345-2773
**Publication License:** 91/13216


**Editorial Office Address:** No.5, Saeedi Alley, Kalej Intersection., Enghelab Ave., Tehran, Iran,
P.O.Box: 13145-799 Tel: (+9821) 88930150 Fax: (+9821) 88930157
E-mail: info@jist.ir , infojist@gmail.com
URL: www.jist.ir


## Indexed by:

| | |
|---|---|
| - SCOPUS | www. Scopus.com |
| - Index Copernicus International | www.indexcopernicus.com |
| - Islamic World Science Citation Center (ISC) | www.isc.gov.ir |
| - Directory of open Access Journals | www.Doaj.org |
| - Scientific Information Database (SID) | www.sid.ir |
| - Regional Information Center for Science and Technology (RICeST) | www.ricest.ac.ir |
| - Iranian Magazines Databases | www.magiran.com |


**Publisher:**
Regional Information Center for Science and Technology **(RICeST)**
Islamic World Science Citation Center **(ISC)**

This Journal is published under scientific support of
Advanced Information Systems (AIS) Research Group and
Digital & Signal Processing Research Group, ICTRC

## Acknowledgement

JIST Editorial-Board would like to gratefully appreciate the following distinguished referees for spending their valuable time and expertise in reviewing the manuscripts and their constructive suggestions, which had a great impact on the enhancement of this issue of the JIST Journal.

**(A-Z)**

- Abdolvand, Neda, Azahra University, Tehran, Iran
- Abdoos, Monire, Shahid Beheshti University, Tehran, Iran
- Aghaie, Abdollah, Khaje Nasir-edin Toosi University of Technology, Tehran, Iran
- Alavi, Seyed Enayatallah, Shahid Chamran University, Ahwaz, Iran
- Ashrafi Peyman, Nosratali, Kharazmi University, Tehran, Iran
- Azhari, Seyed Vahid, Iran University of Science and Technology, Tehran, Iran
- Ghazvini, Mahdieh, Shahid Bahonar University, Kerman, Iran
- Haghighi, Hassan, Shahid Beheshti University, Tehran, Iran
- Kasaei, Shohreh, Sharif University, Tehran, Iran
- khanteymoori, alireza, Zanjan University, Zanjan, Iran
- Koosha, Hamidreza, Ferdowsi University of Mashhad, Mashhad, Iran
- Mavaddati, Samira, University of Mazandaran, Babolsar, Iran
- Mohammadi Zanjireh, Morteza, Imam Khomeini International University, Qazvin, Iran
- Mohammadpur, Davud, Malek Ashtar University of Technology, Tehran, Iran
- Movahedi, Zeinab, Iran University of Science and Technology, Tehran, Iran
- Naderan Tahan, Marjan, Shahid Chamran University, Ahwaz, Iran
- Rezvanian, Alireza, Amirkabir University of Technology, Tehran, Iran
- Roodaki Lavasani, Hoda, Khaje Nasir-edin Toosi University of Technology, Iran
- Sadatrasoul, Seyed Mahdi, Kharazmi University, Tehran, Iran
- Sadegh Mohammadi, Hamidreza, Academic Center for Education Culture and Research (ACECR), Tehran, Iran
- Safkhani, Masumeh, Shahid Rajaee Teacher Training University, Tehran, Iran
- Sarrafzadeh, Abdolhossein, Unitech Institute of Technology, "Auckland", New Zealand
- Shamsinejad, Pirooz, Shiraz University Of Technology, Shiraz, Iran
- Shirvani Moghaddam, Shahram, Shahid Rajaee Teacher Training University, Tehran, Iran
- Shoja, Shamisa, Iran University of Science and Technology, Tehran, Iran
- Yaeghoobi, Kaebeh, Manav Rachna International University, Delhi, India
- Zare, Hadi, Tehran University, Tehran, Iran

# Table of Contents

# A Survey of Two Dominant Low Power and Long Range Communication Technologies

Yas Hosseini Tehrani
Faculty of Electrical Engineering, Sharif University of Technology, Tehran, Iran
yas.hosseini@ee.sharif.ir
Ziba Fazel
Faculty of Electrical Engineering, Sharif University of Technology, Tehran, Iran
ziba_fazel@ee.sharif.ir
Seyed Mojtaba Atarodi*
Faculty of Electrical Engineering, Sharif University of Technology, Tehran, Iran
atarodi@sharif.ir

Received: 24/Feb/2018 Revised: 22/Aug/2018 Accepted: 16/Sep/2018

**Abstract**

The Internet of Things (IoT) connects various kinds of things such as physical devices, vehicles, home appliances, etc. to each other enabling them to exchange data. The IoT also allows objects to be sensed or controlled remotely and results in improved efficiency, accuracy and economic benefits. Therefore, the number of connected devices through IoT is increasing rapidly. Machina Research estimates that the IoT will consist of about 2.6 billion objects by 2020. Different network technologies have been developed to provide connectivity of this large number of devices, like WiFi for cellular-based connections, ZigBee and Bluetooth for indoor connections and Low Power Wide Area Network's (LPWAN) for low power long-distance connections. LPWAN may be used as a private network, or may also be a service offered by a third party, allowing companies to deploy it without investing in gateway technology. Two available leading technologies for LPWAN are narrow-band systems and wide-band plus coding gain systems. In the first one, receiver bandwidth is scaled down to reduce noise seen by the receiver, while in the second one, coding gain is added to the higher rate signal to combat the high receiver noise in a wideband receiver. Both LoRa and NB-IoT standards were developed to improve security, power efficiency, and interoperability for IoT devices. They support bidirectional communication, and both are designed to scale well, from a few devices to millions of devices. LoRa operates in low frequencies, particularly in an unlicensed spectrum, which avoids additional subscription costs in comparison to NB-IoT, but has lower Quality of Service. NB-IoT is designed to function in a 200kHz carrier re-farmed from GSM, with the additional advantage of being able to operate in a shared spectrum with an existing LTE network. But in the other hand, it has lower battery lifetime and capacity. This paper is a survey on both systems. The review includes an in-depth study of their essential parameters such as battery lifetime, capacity, cost, QoS, latency, reliability, and range and presents a comprehensive comparison between them. This paper reviews created testbeds of recent researches over both systems to compare and verify their performance.

**Keywords:** LPWAN; Internet of Things; Narrowband; Wideband; NB-IoT; LoRaWAN.

## 1. Introduction

The Internet of Things (IoT) and its related technologies are predicted to increase expeditiously. According to Machina Research, More than 3.3 billion devices will be connected by 2021 [8]. The Machina research prediction on M2M connections is shown in Fig. 1. The IoT aims at connecting and automating every aspect of our daily life. As shown in Fig. 2, connected devices through IoT, will influence the economy drastically [10]. Therefore, different network technologies have been developed to provide connectivity capable of supporting a large number of devices, which may be located underground, underwater or deep inside buildings. The devices will rely on a wireless connection. Technologies like WiFi based on cellular networks connect devices far from each other, in which power consumption is not limited. For connecting indoor devices which are short distanced with no power limitation, ZigBee, Bluetooth, and similar technologies are appropriate. But in case of restriction over power consumption and battery especially in long distanced communications, Low Power Wide Area Network's (LPWAN) technologies are proposed. LPWAN improves battery life and link budgets, and reduces costs compared to cellular technology [1-4]. For more clarification, a link budget makes a log by keeping all entries of losses and gains in signal propagation. A wave is attenuated via amplifiers and antennas to increase the gain product and eliminate noise. Similarly, data can be lost during propagation of a signal between the transmitter and receiver within one device or between two or more devices. Keeping track of such losses and gains is essential to calculate the reliability and efficiency of a link (through which the transmitter and receiver communicate).

* Corresponding Author

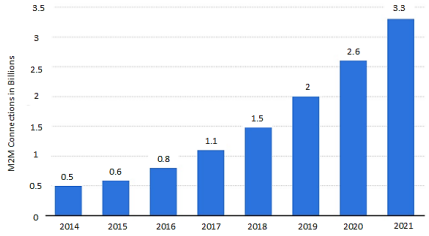

Fig. 1. Billion global connections, 2015- 2021 [8]

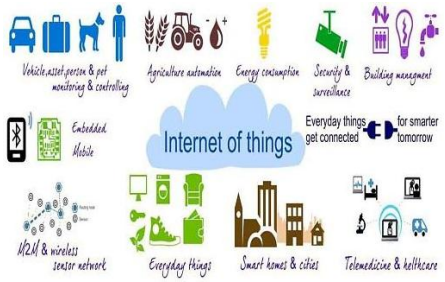

Fig. 2. Internet of Things and Implications in a Developing Economy [10]

As predicted, the IoT devices may rely on LPWAN technologies, which send data over long distance, which enables new types of services. Many technologies like LTE-MTC (LTE Advanced for Machine Type Communications), UBN (ultra-Narrow Band), Senet, Sigfox, Weightless, LoRa, and NB-IoT are supporting new LPWAN approach [4-7]. LPWAN has limitations that need to be discovered clearly. The IoT connectivity technologies segmentation is shown in Fig. 3

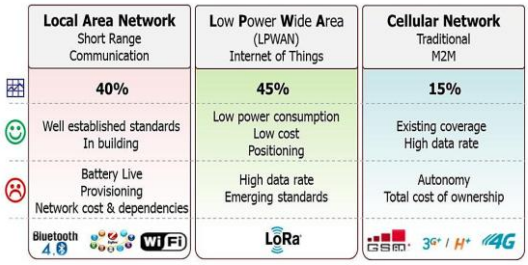

Fig. 3. IoT connectivity technologies segmentation [9]

The goal of this paper is to provide a fair and comprehensive analysis of the capabilities and limitation of LoRaWAN and NB-IoT. The paper is structured as follows: Section 2 provides an overview of the technical description of LoRaWAN and NB-IoT. The critical IoT factors are compared in section 3. Next, the comparisons of measurement results are presented in section 4. Finally, the conclusion is given in section 5.

This research differs from other LPWAN-focused surveys [11-19] in that the scope of LPWA has been broadened to include the most popular, recent and distinct technologies, namely NB-IoT and LoRa, as proprietary solutions, and in that, a more clear and understandable description and a detailed direct comparison of them have been performed.

## 2. Technical Description of LoRaWAN & NB-IoT

Two available leading technologies for LPWAN, i.e., LoRaWAN and NB-IoT are described technically in this section. In the first one, coding gain is added to the higher rate signal to combat the high receiver noise in a wideband receiver, while receiver bandwidth is scaled down to reduce noise seen by the receiver in the second one. Different methods cause different functionality and specification for each one, which is explained in the following.

### 2.1 LoRa & LoRaWAN

The LoRa is a newborn technology in recent years. It can operate in non-licensed sub-1GHz frequency bands, i.e., frequency bands from 400MHz to 900MHz. Therefore, it has region specific configuration problems.

LoRa consists of two primary layers: a physical layer and a MAC layer protocol (LoRaWAN). The physical layer is based on the spread spectrum modulation scheme. An increased link budget, as well as better immunity to network interference, is achieved by deploying a derivate of chirp spread spectrum modulation (CSS) [2]. LoRa allows usage of configurable bandwidth of 125kHz, 250kHz, or 500kHz. Larger bandwidths support higher data rate, shorter time on air, but lower sensitivity. Therefore, as the bandwidth becomes wider, the resistance to channel noise, Doppler effects, long-term relative frequency and fading will increase [2,8].

The transmitter generates chirp signals by varying their frequency over time and keeping phase between adjacent symbols constant. A time domain equation of single chirp waveform is presented in (1).where$\Phi(t)$ is the phase of chirp waveform.

$$c(t) = \begin{cases} \exp(j\Phi(t)) & -\frac{T}{2} \le t \le \frac{T}{2} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

LoRa modulation depends on

- Coding rate (CR), a measure of the amount of forwarding error correction;
- Spreading factor (SF), a ratio between the chip rate and the underlying the symbol rate (7-12);
- Bandwidth (BW), the frequency interval (125kHz-500kHz).

The LoRaWAN specification of different countries is summarized in Table. 1. The communication going from an antenna to nodes is called downlink, and when it is going from a node to an antenna is called uplink.

Table 1. LoRaWAN specification of different countries [9].

| | ***Europe*** | ***North America*** | ***China*** | ***Korea*** | ***Japan*** |
|---|---|---|---|---|---|
| ***Frequency Band*** | 867-869MHz | 902-928MHz | 470-510MHz | 920-925MHz | 865-867MHz |
| ***Number of channels*** | 10 | 64+8+8 | In definition by Technical Committee | | |
| ***Channel BW Uplink*** | 125/250kHz | 125/500kHz | | | |
| ***TX Power Uplink*** | 125kHz | 500kHz | | | |
| ***TX Power Downlink*** | +14dBm | +20dBm | | | |
| ***TX power Downlink*** | +14dBm | +27dBm | | | |
| ***SF Uplink*** | 7-12 | 7-10 | | | |
| ***Data rate*** | 250bps- 50kbps | 980bps- 21.9kbps | | | |
| ***Link Budget Uplink*** | 155dB | 154dB | | | |
| ***Link Budget Downlink*** | 155dB | 157dB | | | |

LoRaWAN is the MAC layer above the LoRa. The LoRaWAN's architecture is based on a star topology.

Multiple LoRa End devices are connected to Gateways. In Europe, LoRaWAN's are limited to 10 channels, has duty cycle restrictions, without channel time limitations. LoRa WANs in North America have 64 channels. LoRaWAN network layers are shown in Fig. 4.

LoRaWAN supports three different classes. Class A must be supported by all end devices, and it has the lowest power consumption [2,9].

- Class A of end devices allows bi-directional communications. An uplink transmission is followed by two downlinks receive windows.
- Class B of end devices opens other receive windows at scheduled times.
- Class C of end devices has continuously- open receive windows.

LoRaWAN communication profile classes are shown in Fig. 5.

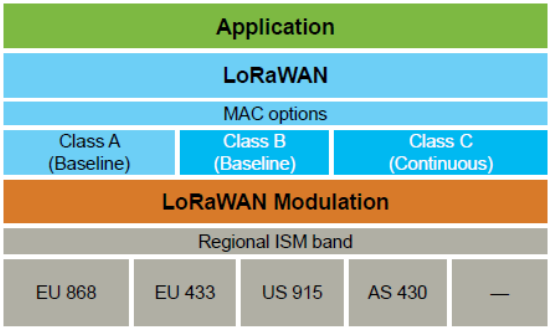


Fig. 4. LoRaWAN network layers [9]

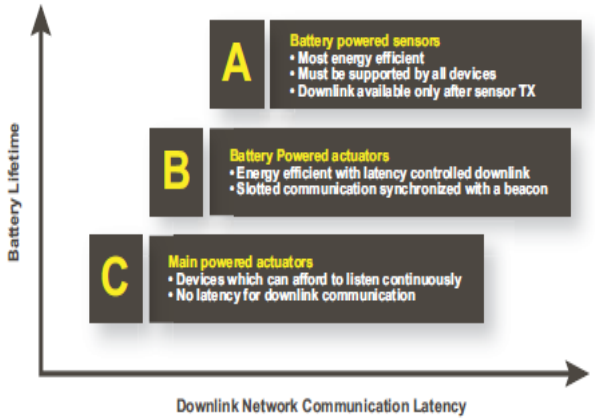


Fig. 5. LoRaWAN communication profile classes [9]

## 2.2 NB-IoT

NB-IoT is a simple subset of long-term evolution (LTE) standards suitable for IoT. NB-IoT does not have many features of LTE such as dual connectivity, channel quality measurement, etc. to be simple, chip and low power, which is a necessity for IoT. It uses sub-1GHz licensed frequency bands, i.e., frequency bands from 400MHz to 900MHz and employs QPSK modulation [1]. Narrowband modulation techniques encode the signal in a narrow bandwidth and share the overall spectrum very efficiently between multiple links. Moreover, it lowers the noise level inside a single narrowband and hence provides a high link budget. This modulation scheme needs no processing gain, resulting in simple and inexpensive transceiver design. The architecture of NB-IoT network is shown in Fig. 6.

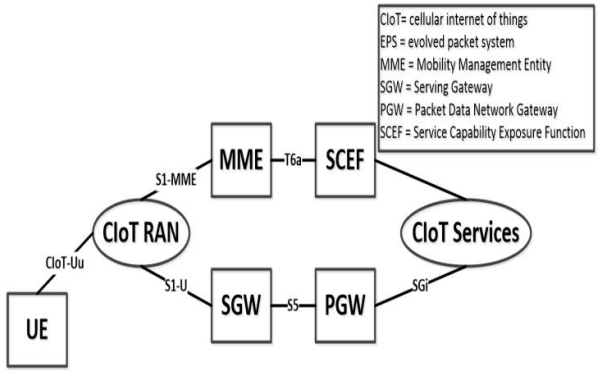


Fig. 6. NB-IoT network architecture [10]

As expected, NB-IoT core network is based on the evolved packet system (EPS) similar to LTE. Two optimizations for the cellular IoT are additionally defined, which are the user plane optimization and the control plane optimization [2]. Both planes choose the best path for user and control data packets, for uplink and downlink data. The cell access procedure of an NB-IoT user is also similar to that of LTE. Therefore, as LTE is already widespread in the US, IoT is generally based in NB-IoT there [10]. Fig.7 shows message flow for Random Access Channel (RACH) procedure Utilized by NB-IoT.

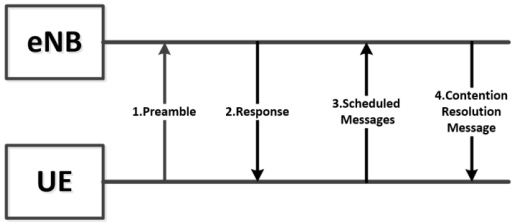


Fig. 7. NB-IoT message flow for RACH procedure [10]

## 3. Comparison of IoT Factors

Many factors should be considered to choose the suitable technology for an IoT application, such as quality of service (QoS), battery lifetime, latency, capacity, deployment model, coverage, range, cost, and security. These factors are discussed in this section based on section 2 and Table 2, which summarizes some features of NB-IoT and LoRaWAN.

Table 2. Some Features of LoRaWAN and NB-IoT [1, 4-7].

| ***Parameter*** | ***LoRaWAN*** | ***NB-IoT*** |
|---|---|---|
| Spectrum | Unlicensed | Licensed |
| Modulation | Chirp Spread Spectrum | QPSK |
| Bandwidth(kHz) | 125-500 | 25-180 |
| Peak data rate in up-link(kb/s) | 290-50 | 204.8 |
| Latency(s) | Multiple of 10 | <10 |
| Range(km) | <15 | <35 |
| Power efficiency | High | Medium |
| Deployment model | Network operators and small companies | Network operators |
| Cost | Medium | High |
| Security | Medium to high | High |

### 3.1 QoS

Although LoRaWAN uses chirp spread spectrum modulation to avoid interferences, noise, multipath fading and shadowing phenomena; it cannot provide the best of service quality due to the unlicensed frequency band and asynchronous protocol utilization. On the contrary, NB-IoT uses licensed frequency band, and its synchronous timing windows protocol is optimized for best possible QoS. Worth to mention that NB-IoT has more cost compared to LoRaWAN, to provide improved QoS [1].

### 3.2 Battery Lifetime and Latency

End nodes in LoRaWAN are asynchronous and only communicate whenever they have some data to send to gateways. This kind of protocol is called "ALOHA." NB-IoT uses a synchronous protocol in which, nodes are periodically turned on to check whether they have new data to send or not. This synchronizing consumes a significant portion of power and shortens the battery lifetime. Therefore, LoRaWAN consumes less power. Some neglect this good aspect only due to its little more latency caused by the long process of demodulation, which is not correct. In other words, comparison revels that for high data rate and low latency applications, NB-IoT is a better choice while LoRaWAN suits applications in need of longer battery lifetime but not soon data arrival more [8,9].

### 3.3 Capacity

Some may misunderstand and conclude waste of frequency band due to LoRa modulation rather than NB-IoT. They must pay attention to another fact that LoRa utilizes different spread spectrums to send some signals through a channel simultaneously, which increases capacity. For more clarification, assume a narrow band system with the bandwidth of 125 KHz with a rate of 1.2 Kb/s. In the first scenario (i.e., NB-IoT communication), 12 channels of narrow band width FSK modulation with the data rate of 1.2 Kb/s results in the capacity of 14.4 kb/s according to equation 2.

$$\text{Capacity}_{\text{ch}} = \text{Number}_{\text{ch}} \times \text{BR}_{\text{ch}} \quad (2)$$

Where BR and ch stand for bit rate and channel, respectively.

While in the second scenario (i.e., LoRa communication), still with the usage of same bandwidth, the capacity of only one channel will be greater taking advantage of different spread spectrums. The capacity is calculated according to equation 3 and equals $21.531\text{Kb/s}$.

$$\text{Capacity}_{\text{ch}} = \text{Number}_{\text{ch}} \times (\text{BR of SF12:6}) \quad (3)$$

Where BR of SF 12:6 stands for the bit rate of each spreading factor from SF=12 down to SF=6. These bit rates are equal to 293, 573,976, 1757, 3125, 5468 and 9375 b/s respectively.

As a result, LoRa modulation can increase channel capacity up to 50% [2].

### 3.4 Deployment Model

NB-IoT is a subsection of LTE born in June of 2016. Hence, its networks can be deployed by adapting and reusing already available cellular networks; it means it is available wherever the cellular network is available and some time is needed for adaption of networks before successful deployment. Another side of the coin, LoRaWAN's ecosystem is matured and ready to be either deployed by large companies of cellular networks or small start-up companies. The LoRaWAN network architectures simpler, but the network server is more complex.

### 3.5 Coverage and Range

The most important advantage of LoRaWAN is its ability to cover a whole city with only one gateway or base station; for example, all around Belgium with an area of 3500Km2 is covered by only seven LoRaWAN base stations [8]. In contrary, NB-IoT is deployable only where 4G/LTE base stations are available which means it could not cover rural and the country regions. Still, NB-IoT has a wider range than LoRaWAN.

### 3.6 Cost

Cost is a summation of spent money on different parts such as frequency band, network, device, and deployment. For example, LoRaWAN pays less money for each gateway than NB-IoT. It does not pay any for frequency band either. However, its device is more expensive than NB-IoT's [1]. In total, LoRaWAN is less expensive.

### 3.7 Security

As NB-IoT adapts and reutilizes the available cellular networks, it does gain their security too. NB-IoT's standard protocol is half-close due to its network and half-open due to the open access device. In contrary, LoRaWAN's physical layer is open to all people, small companies, and large network companies which results in its unsecured inherent; therefore, to solve his issue, two layers of security to protect data of users are defined.

### 3.8 Gateway

The dedicated gateways are necessary for LoRa to function correctly, while NB-IoT eliminates the need for the same. So the LoRa gate ways can be a potentially extra problem. In NB-IoT, these are not required.

### 3.9 Operability in Private Networks

LoRa can be used by private companies in their proprietary networks, but NB-IoT can't be. NB-IoT can only be a user in public models.

### 3.10 Modulation & Complexity

LoRa has a specific modulation method. This modulation method is based on spread spectrum to support long range, but in the other hand, this method have lower data rate in comparison with NB-IoT modulation method. Also, there are issues about IP rights and licensing of LoRa technology. NB-IoT technology

uses DSSS modulation and has lower hardware complexity in comparison to LoRa.

The direct comparison of LoRa and NB-IoT Technologies is illustrated in Table 3. From the results, we can conclude that LoRa hardware can be produced and sold at a low price, and LoRa devices have no subscription cost, but the LoRa Alliance only has limited control over the deployment of networks. NB-IoT technologies on the other hand, though the devices will have a subscription charge, the deployment of gateways is for newer grade hardware as simple as applying a software update; a country can have a functioning NB-IoT network within an hour.

Table 3. Direct Comparison of LoRa & NB-IoT Technologies.

| ***Technology*** | ***LoRa*** | ***NB-IoT*** |
|---|---|---|
| Topologies supported | Typically Star, Mesh possible | Star |
| Maturity Level | Early stages- some deployment | Early Stages |
| Frequency Band | Sub GHz ISM bands | LTE & GSM bands |
| MAC Layer | ALOHA-based | LTE-based |
| Founded | 2015 | 2016 |
| Modulation Technique | Spread Spectrum | LTE-based |
| Proprietary aspects | Physical layer | Full Stack |
| Nodes per gateway | >1,000,000 | 52,000 |
| Deployment model | Private and Operator -based | Operator-based |
| Encryption | AES | 3GPP |
| Interference immunity | Very High | Low |
| Energy efficiency | Better than NB-IoT | good |

## 4. Comparison of Measurement Results

In this section, we compare the measurement results of recent works on IoT factors. The target is to study the important factor of LoRa WAN and NB-IoT technologies in a real condition. [20] studied NB-IoT coverage and capacity for a small country side area. Also, coverage and capacity have been studied for NB-IoT [21] and LoRa [20]. In [22] the coverage of NB-IoT, LoRa, GPRS, and Sigfox has been simulated in a realistic scenario, covering $7800\,km^2$, using Telenor's commercial 2G, 3G, and 4G deployment. According to the reported measurement results, the NB-IoT is the best performing indoor solution. It has less than a 4% failure rate for ten devices, while LoRa provides 20% failure rates, which also has more sensitivity to device numbers. Therefore, NB-IoT has the best coverage and link adoption, but it has the longer time on air.

The LoRa has only one manufacturer (Semtech). Also, as the unlicensed bands become more crowded, the interference may increase. Therefore, the performance of the LoRa networks may decrease.

According to the experiment's results in [23], the lost packet number depends on SF, CR, and BW of the communication in the presence of different noise levels. The transmitter was placed 10 meters from the receiver. By changing Bandwidth from 125 kHz to 500 kHz, when the SNR is -10dBm, the packet lost increased from 1% to 30.66%. The measurement results of LoRa in recent literature are summarized in Table 4.

From the results of the recent works, summarized in Table 4, we can conclude LoRa features low data rate and long communication range. Therefore it is suitable for applications with a reduced number of messages without challenging delay constraints. While it can't be an appropriate solution for real-time applications which require low latency.

The coverage analysis has been done for NB-IoT and LoRa in [36]. Two different configurations in 800 MHz with 25 and 500 active sectors have been simulated to carry out their maximum connectivity level. The results show that LoRa covered 80% area, while NB-IoT can reach 90% of the covered area, because, NB-IoT is a licensed solution.

In [37], capacity and system efficiency analyses are performed for a massive NB-IoT system for smart metering. The simulations have done in a realistic scenario, in a rectangular area of 2000m ×1700m. In this configuration, 75% of transmitting uplinks are succeeded. And the maximum measured coverage distance is about 960m for a single receiver and transmitter configuration.

## 5. Conclusion

Wireless communication is the most recent industrial revolution in the last decades which is utilized for connecting devices to each other. More than twenty-five billion machines and objects which are considered as devices are expected to be connected by the year 2020. There are many challenges and factors such as connection range, data rate, power, etc. to consider to provide connectivity of these devices. Various network technologies such as local area network, LPWAN, and Cellular network is currently available. This paper has focused on the most leading wireless technology for long range and low power communication, i.e., LPWAN.

In this work, two prominent LPWAN technologies, i.e., LoRaWAN and NB-IoT are described, analyzed, and compared in depth. LoRaWAN is a coding gain method which defeats the noise of long range by a new method of modulation in its unlicensed frequency band, while NB-IoT does so by utilizing minimum possible licensed frequency band. It is shown that licensed NB-IoT has the advantage of better QoS, latency, and range, while unlicensed LoRaWAN has advantages of better battery lifetime, capacity, and cost. Therefore, the choice is strongly depending on the users' goals and necessities of each application.

Table 4. Measurement results of LoRa in recent literature.

| ***Ref*** | ***#Gate ways*** | ***#Nodes*** | ***BW (kHz)*** | ***Tx (dBm)*** | ***SF*** | ***ISM band (MHz)*** | ***RSSI measurement*** | ***Payload (byte)*** | ***Coverage (km)*** | ***Reliability measurement*** |
|---|---|---|---|---|---|---|---|---|---|---|
| [22] | 1 | 10 | 125 | 2,4,6,8,10,12 | - | 868 | Done | - | $7800km^2$ area | Maximum Coupling Loss (MCL) |
| [23] | | | 125,250,500 | | 7,8,9,10,11 | | | 1 | 10m with the presence of with Gaussian | % packets lost |

| *Ref* | *#Gate ways* | *#Nodes* | *BW (kHz)* | *Tx (dBm)* | *SF* | *ISM band (MHz)* | *RSSI measurement* | *Payload (byte)* | *Coverage (km)* | *Reliability measurement* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | noise | |
| [24] | 1 | 1 | 125,250,500 | | 7,8,9,10,12 | 868 | | 106 | | Packets PRR |
| [25] | 1 | 1 | 125 | 20 | 12 | - | Done | large | 7 floors | RSSI value |
| [26] | 3 | 2 | - | 14 | variable | 868 | Done | seq. num | 2.2 | record ACKs |
| [27] | 0 | 6 | variable | 17 | variable | - | - | variable | 0.342 | packet reception rate |
| [28] | 1 | 1 | - | - | variable | - | - | - | 2 | SF for coverage |
| [29] | 1 | 1 | 125 | 14 | variable | 868 | Done | variable | 0.5m-60m | # packets lost & packets error |
| [30] | 1 | 1 | 250 | - | 10 | 868 | Done | 10,50,100 | 0.276-8.52 | PER |
| [31] | 0 | 7 | - | - | - | 915 | Done | 26 | 0.5-2.7 | % valid packets received |
| [32] | 1 | 1 | 125 | 14 | 12 | 868 | Done | seq. num | 15 | % packets lost |
| [33] | 1 | 1 | 125 | 14 | 12 | 868 | - | - | 65m-195m | %received packets |
| [34] | 0 | 2 | - | 3 | variable | 2450 | - | 21 | 0.975 outdoor 30 m indoor | % valid packets |
| [35] | 1 | 1 | 125 | 2/14 | variable | - | Done | 1,25,51 | 3.4 outdoor 100m indoor | % packets received & avg.throughput |

## References


[1] R. Sinha, Y. Wei, and S. Hwang; "A survey on LPWA technology: LoRa and NB-IoT," Elsevier, ICT Express 3, 14–21, 2017.

[2] Semtech, AN 120022, LoRa Modulation Basics, May 2015. Available: http://www.semtech.com/images/datasheet/an1200.22pdf.

[3] Link Labs, "NB-IoT vs. LoRa vs. Sigfox," Home Blog, January 23, 2017.

[4] H. Wang and A. O. Fapojuwo, "A Survey of Enabling Technologies of Low Power and Long Range Machine-to-Machine Communications," IEEE Communications Surveys & Tutorials, DOI 10.1109/COMST.2017.2721379.

[5] INGENU, "RPMA Technology for the Internet of Things," 14 March 2017. Available: http://theinternetofthings.report/Resources/Whitepapers/4cbc5e5e-6ef8-4455-b8cd-f6e3888624cb_RPMA%20Technology.pdf.

[6] Nokia, "LTE evolution for IoT connectivity," Nokia, 2016. Available: http://resources.alcatel-lucent.com/asset/200178.

[7] R. Ratasuk, N. Mangalvedhe, Y. Zhang, M. Robert and J. P. Koskinen, "Overview of narrowband IoT in LTE Rel-13," 2016 IEEE Conference on Standards for Communications and Networking (CSCN), Berlin, 2016.

[8] Nicolas Ducrot et al., Olivier Hersent et al.; "LoRa Device Developer Guide," Orange Connected Objects & Partnerships plus Activity, April, 2016.

[9] LoRa Alliance, "A technical overview of LoRa and LoRaWAN," Technical Marketing Workgroup 1.0, 2016.

[10] D. Rohde, J. Schwarz, Narrowband Internet of Things, Aug., 2016.Available: https://www.rohdeschwarz.com/us/applications/narrowband-Internet-of-things-application-note 56280-314242.html

[11] F. Samie, L. Bauer, and J. Henkel. "IoT Technologies for Embedded Computing: A Survey." Proceedings of the Eleventh IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis. CODES '16. 2016, 8:1–8:10.

[12] S. Andreev et al. "Understanding the IoT connectivity landscape: a contemporary M2M radio technology roadmap" IEEE Communications Magazine, 2015, pp. 32–40.

[13] M. Elkhodr, S. Shahrestani, and H. Cheung. "Emerging Wireless Technologies in the Internet of Things: A Comparative Study." International Journal of Wireless & Mobile Networks (IJWMN), 2016.

[14] U. Raza, P. Kulkarni, and M. Sooriyabandara. "Low Power Wide Area Networks: An Overview," IEEE Communications Surveys Tutorials, 2017.

[15] R. Sanchez-Iborra and M. Cano. "State of the Art in LP-WAN Solutions for Industrial IoT Services." Sensors, 2016.

[16] A. Ali et al. "Technologies and challenges in developing Machine-to-Machine applications: A survey." Journal of Network and Computer Applications, 2017, pp. 124 –139.

[17] B. Moyer. "Low Power, Wide Area: A Survey of Longer-Range IoT Wireless Protocols." EE Journal, 2015.

[18] Q. Song, L. Nuaymi, and X. Lagrange. "Survey of radio resource management issues and proposals for energy efficient cellular networks that will cover billions of machines." EURASIP Journal on Wireless Communications and Networking, 2016.

[19] W. Guibene, K. E. Nolan, and M. Y. Kelly. "Survey on Clean Slate Cellular-IoT Standard Proposals." 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, 2015.

[20] K. Mikhaylov, J. Petaejaejaervi, and T. Haenninen, "Analysis of Capacity and Scalability of the LoRa Low Power Wide Area Network Technology," European Wireless, May 2016, pp. 1–6

[21] N. Mangalvedhe, R. Ratasuk, and A. Ghosh, "NB-IoT Deployment Study for Low Power Wide Area Cellular IoT," PIMRC, 2016.

[22] M. Lauridsen, H, Nguyen, B. Vejlgaard, I. Kovacs, and P. Mogensen, "comparison of GPRS, NB-IoT, LoRa, and SigFox in a 7900 km² area", Vehicular Technology Conference (VTC Spring), 2017 IEEE 85th, Nov. 2017.

[23] L.Angrisani, P.Arpaia, F. Bonavolonta, M. Conti, and A. Liccardo, "LoRa Protocol Performance Assessment in Critical Noise Conditions," Research and Technologies for Society and Industry (RTSI), 2017 IEEE International Forum on, Sept. 2017.

[24] C.Orfanidis, L.Feeney, M. Jacobsson, and P. Gunningberg, "Investigating interference between LoRa and IEEE

802.15.4g networks", 2017 IEEE 13th international Conference on Wireless and Mobile Computing, Network and Communications (WiMOB), 2017.
[25] L. Gregora, L. Vojtech, and M. Neruda, "Indoor Signal Propagation of LoRa Technology," 2016 17th International Conference on Mechatronics Mechatronika (ME), Prague, Czech Republic, 2016, pp. 1–4.
[26] A. Wixted, P. Kinnaird, A. Tait, A. Ahmadinia, and N. Strachan, "Evaluation of LoRa and LoRaWAN for Wireless Sensor Networks," 2016 IEEE SENSORS, October 2016, pp. 1–3.
[27] M. Bor, J. Vidler, and U. Roedig, "LoRa for the Internet of Things," Proceedings of the 2016 International Conference on Embedded Wireless Systems and Networks, Graz, Austria, February 2016, pp. 361–366.
[28] A. Zanella and M. Zorzi, "Long-Range Communications in Unlicensed Bands: The Rising Stars in the IoT and Smart City Scenarios," IEEE Wireless Communications, vol. 23, no. 5, pp. 60–67, October 2016.
[29] P. Neumann, J. Montavont, and T. No¨el, "Indoor Deployment of Low-Power Wide Area Networks (LPWAN): a LoRaWAN case study", 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), New York, NY, USA, October 2016, pp. 1–8.
[30] M. Aref and A. Sikora, "Free-space range measurements with Semtech LoRaTM technology," 2014 2nd International Symposium on Wireless Systems within the Conferences on Intelligent Data Acquisition and Advanced Computing System (IDAACS-SWS 2014), Offenburg, Germany, September 2014, pp. 19–23.
[31] Lukas, W. A. Tanumihardja, and E. Gunawan, "On the application of IoT: Monitoring of troughs water level using WSN," 2015 IEEE Conference on Wireless Sensors, ICWiSE 2015, Melaka, Malaysia, August 2016, pp. 58–62.
[32] J. Pet¨aj¨aj¨arvi, K. Mikhaylov, A. Roivainen, T. H¨anninen, and M. Pettissalo, "On the coverage of LPWANs: Range evaluation and channel attenuation model for LoRa technology," 2015 14th International Conference on ITS Telecommunications (ITST 2015), Copenhagen, Denmark, December 2016, pp. 55–59.
[33] J. Pet¨aj¨aj¨arvi, K. Mikhaylov, M. H¨am¨al¨ainen, and J. Iinatti, "Evaluation of LoRa LPWAN technology for remote health and wellbeing monitoring," International Symposium on Medical Information and Communication Technology (ISMICT), Worcester, MA, USA, March 2016, pp. 1–5.
[34] T. Wendt, F. Volk, and E. Mackensen, "A benchmark survey of Long Range (LoRa TM) Spread Spectrum-Communication at 2.45 GHz for safety applications", Wireless and Microwave Technology Conference (WAMICON), 2015 IEEE 16th Annual, Cocoa Beach, FL, USA, April 2015, pp. 1–4.
[35] A. Augustin, J. Yi, T. Clausen, and W. Townsley, "A Study of LoRa: Long Range & Low Power Networks for the Internet of Things," Sensors, vol. 16, no. 9, 2016.
[36] S. persia, C. Carciofi, and M. Faccili, "NB-IoT and LoRa Connectivity analysis for M2M/IoT Smart Grid Applications," 2017 AEIT International Annual Conference, Dec. 2017.
[37] M.Pennacchiono, M. Gabriella, T, Oercorella, C. Carrlini, "NB-IoT System Deployment for Smart Metering: Evaluation of Coverage and Capacity Performances," 2017 AEIT International Annual Conference, Dec. 2017.

**Yas Hosseini Tehrani** received the B.Sc. degree in Electrical engineering from Tehran University, Tehran, Iran, in 2014. She received M.Sc. degree in Electrical engineering from Tehran University, Tehran, Iran, in 2016. She is currently Ph.D. candidate in Department of Electrical Engineering of Sharif University, Tehran, Iran. Her area research interests include Ultra low power system design, Instrumentation, RF, analog, and mixed signal system design, Internet of things, Switched-mode power supply, and Parallel Programming.

**Ziba Fazel** was born in Urmia, Iran in 1990. She received the B.Sc. degree in electrical engineering from University of Tehran, Tehran, Iran in 2012, and the M.Sc. degree in electronics engineering from Sharif University of Technology (SUT), Tehran, Iran, in 2015. She is currently pursuing her Ph.D. in Electronics engineering at SUT. She has been a Teaching Assistant for (under)grad courses at Electrical Engineering (EE) Department, SUT since 2015. From 2016, she is an Analog/Mixed-signal Design Engineer with Sharif Integrated Circuits and Systems Group at Electrical Engineering (EE) department, SUT. Her main research interests include RF/analog/mixed-signal integrated systems and circuits for low power wireless communications and electronic interfaces of different daily life applications.

**Seyed Mojtaba Atarodi** received his Ph.D. degree from the University of Southern California (USC), Los Angeles, CA, USA, on the subject of analog IC design in 1993. He received the M.Sc. degree in electrical engineering from the University of California, Irvine, CA, USA, in 1987 and B.S.E.E. from Amir Kabir University of Technology (Tehran Polytechnic), Iran, in 1985. From 1993 to 1996, he worked with Linear Technology Corporation as a Senior Analog Design Engineer and produced 2 IC products in the field of high frequency high dynamic range continuous-time Gm-C filters. Since then, he has been consulting with different IC companies. He is currently an Associate Professor at Sharif University of Technology, Tehran. He has published more than 100 technical journal and conference papers in the area of analog/RF and mixed-signal integrated circuit design. He is the author of two books on analog CMOS IC design and Integrated Filter design from DC to RF. His main research interests are integrated bioelectronics, RF/analog/mixed-signal ICs, and analog CAD tools. Dr. Atarodi is a senior member of IEEE and the recipient of Interntional Kharazmi Award, the most prestigious knowledge-based industrial award in Iran.

# A Novel User-Centric Method for Graph Summarization Based on Syntactical and Semantical Attributes

Nosratali Ashrafi Payaman
Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
ashrafi@khu.ac.ir
Mohammad Reza Kangavrai*
Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
kangavari@iust.ac.ir

Received: 25/04/2018 Revised: 05/09/2018 Accepted: 28/10/2018

**Abstract**

In this paper, we proposed an interactive knowledge-based method for graph summarization. Due to the interactive nature of this method, the user can decide to stop or continue summarization process at any step based on the summary graph. The proposed method is a general one that covers three kinds of graph summarization called structural, attribute-based, and structural/attribute-based summarization. In summarization based on both structure and vertex attributes, the contributions of syntactical and semantical attributes, as well as the importance degrees of attributes are variable and could be specified by the user. We also proposed a new criterion based on density and entropy to assess the quality of a hybrid summary. For the purpose of evaluation, we generated a synthetic graph with 1000 nodes and 2500 edges and extracted the overall features of the graph using the Gephi tool and a developed application in Java. Finally, we generated summaries of different sizes and values for the structure contribution (α parameter). We calculated the values of density and entropy for each summary to assess their qualities based on the proposed criterion. The experimental results show that the proposed criterion causes to generate a summary with better quality.

**Keywords:** Graph Summarization; Summary Graph; Super-node; Semantical Summarization.

## 1. Introduction

Graph is widely used for modeling data and their relationships. Social networks, communication networks, web graphs, biological networks, and chemical compounds are examples of data modeling by graphs. There are several applications that generate large scale and massive graphs and extensive research has been undertaken about the theory and engineering of terra-scale graphs [1]. These graphs are massive with a high growth rate. To clarify the issue, consider the statistics of Facebook users [2], which increased from one million at the end of 2004 to 1.11 billion in March 2013.

Recently proposed graph summarization algorithms [3]-[7] reduce a massive graph to a smaller one by removing details and preserving general properties. The smaller graph can be used for query answering. Of-course, these answers are not exact and may be some errors. This kind of error is acceptable since the queries are responded quickly, required by many applications.

The most real-life applications generate attributed graphs, and a summary based on both structure and vertex attributes is a critical requirement in these applications. Structural and attribute-based relationship of vertices can be considered as current and future relationships of vertices, respectively. The first one is obvious and the second one can be justified in that, based on the existing statistics of social networks, the vertices with common attribute values are probably connected. Thus generating a summary based on both structure and vertex attributes has received growing attention of the computer scientists recent years.

Although generating attribute-based summaries is not difficult and several algorithms [8] have been proposed for this purpose, generating a summary based on both structure and vertex attributes (hybrid summary) with user-determining the degrees of each is nothing less of challenge. It is obvious that the importance of structure and vertex attributes for summarization is not the same in all applications and therefore it is reasonable to consider variable weighting coefficients for them. Recently two algorithms [9],[10] have been proposed for hybrid summarization and clustering.

In graph summarization, ontology is a critical component, required to generate a high quality summary. Ontology helps to a summary in fitting with a user's needs and appropriate size. Using ontology, it is possible to explore the relationship between two attributes, determining whether they are the same, different, one subtype of each other, among other things. By involving ontology in summarization process, it is possible to drill down or roll up on the resultant summary and generate a summary of the right size. The previous summarization methods, discussed in this paper, do not incorporate ontology in the summarization process. To the best of our knowledge, no previous method is capable of generating a summary graph that can interact with both the knowledge base and the user.

In this paper, an ontology-based interactive method has been proposed for graph summarization. This method

* Corresponding Author

can be used for various types of summarization such as structural, attribute-based or hybrid.

The proposed method has a number of advantages such as generality, user-centric, knowledge-based and interactiveness nature, which makes it ideal for graph summarization. In the following, some of these advantages have been presented.

**Generality:** By specifying the similarity measure of two vertices, the proposed method can generate any kind of summary including structural, attribute-based or hybrid ones.

**User-centric**: The proposed method can generate a summary based on structure, vertex attributes or both. The importance of the structure and vertex attributes in summarization and also the significance of the attributes can be determined by the user and incorporated in the summarization.

**Knowledge-based**: Using a knowledge base leads to the generation of a summary of appropriate size and that is consistent with user's needs.

**User-interactive**: The summarization process is a supervised method in which the user can decide to stop summarization through interacting with the program.

**The proposed criterion**: We propose a new criterion for termination of the summarization process. We defined the proposed criterion based on the density and entropy, which shows the quality of the hybrid summary.

The rest of this paper is organized as follows. In Section 2, related works are reviewed. Section 3 is dedicated to graph summarization and related definitions. Section 4 presents the proposed method for graph summarization. The experimental results are provided in Section 5. Discussions are presented in Section 6 and finally conclusions are drawn in Section 7.

## 2. Related Works

In this section, we review previous works on three different types of graph summarization to discuss the main challenges of graph summarization.

### 2.1 Structural Summarization

Navlakha et al. [3] proposed a summarization algorithm for structural summarization where graph compression was performed by partitioning similar nodes into one group and dissimilar nodes into different groups. Edges between every pair of super-nodes are aggregated and constitute a super-edge in the summary graph. In this method, a graph is compressed with the minimum representation cost based on the MDL[1] idea. At first, they developed a GREEDY algorithm for this purpose and then proposed a RANDOMIZED version to reduce the run-time.

In [11] another method has been proposed to summarize structural graphs. In this method, the quality of a summary is guaranteed and the graph is summarized with the aim of minimizing the reconstruction errors. The authors of this paper have presented a connection between graph summarization and geometric clustering. Based on this connection, they developed a polynomial-time algorithm to compute the best possible summary of a certain size.

In [12], three distributed algorithms have been proposed to summarize large scale graphs. These algorithms are DistGreedy, DistRandom and DistLSH which differ in how they select a pair of nodes for merging (greedy, randomly, and using locality sensitive hashing theory, respectively).

Structural summarization can be used for mining frequent patterns. Chen et al. [13] proposed a method for identifying frequent patterns by producing randomized summary graphs. In fact, summary graphs rather than original graphs are mined, which are massive and time consuming. Graph summarization can also be beneficial for subgraph mining [14] and classification of aid flyers based on their property types [15].

Structural summarization can be performed using spectral graph clustering that partitions a graph based on eigenvalues and eigenvectors of the graph adjacency matrix [16]-[20]. This technique is widely used in image segmentation and social network analysis. Spectral clustering has also extensive applications in finding communities in networks [21]. It initially converts a large graph into a small one by summarization and then the resultant small summary graph is clustered by spectral clustering [22].

Graph summarization also is also used in community detection and a growing body of literature [23]-[27] has been published on this subject.

### 2.2 Attribute-Based Summarization

In [8], a summarization method with two novel operations has been proposed. These operations are called SNAP[2] and k-SNAP for which used for grouping nodes and summarizing attributed graphs. Attribute compatible grouping and relation compatible grouping defined by the authors of this paper. They also improved SNAP operation by proposing k-SNAP operation, where k was the summary size determined by the user.

In 2009, Zhang et al. [5] improved the k-SNAP operation by proposing the CANAL algorithm, to categorize attribute values automatically, and providing a criterion to measure the quality of a summary.

In 2008, Chen et al. [28] proposed the OLAP framework. In this framework, the cubes were created on the graph based on dimensions and measures. In the OLAP framework, a graph is summarized based on the selected attributes and input information.

### 2.3 Hybrid Summarization

For clustering a graph based on both structure and vertex attributes, a method was proposed in [10]. In this method, a new graph with real and virtual links is constructed for a given graph. The virtual links are added to the new graph on the account of attribute-based similarity of vertices. This new constructed graph is called the augmented graph. The similarity of two nodes is measured based on both real and virtual links in the augmented graph.

1. Minimum Description Length

2. **S**ummarization by Grouping **N**odes on **A**ttributes and **P**airwise Relations

Another method of hybrid summarization was proposed in [9]. This method first summarizes a graph based on attributes and then adjusts the summary to the graph structure by moving nodes between super-nodes.

The main challenge in the graph summarization methods is the absence of an ontology-based method to generate a hybrid summary of the attributed graph in which the α is specified by the user. The methods proposed in [9],[10] are not ontology-based and therefore unsuitable for this purpose. The proposed method resolves these challenges.

## 3. Graph Summarization Notion

We here present symbols and abbreviations that, have been used in Section 3.1. In Sections 3.2, 3.3, and 3.4 we will illustrate the concept of structural, attribute-based and hybrid summarization. We also present definitions of graph summarization in these sections.

### 3.1 Notations

In this section, the most frequently used symbols and abbreviations in this paper have been listed in Table 1.

### 3.2 Structural Summarization

The definition of a summary graph according to [5] is as follows:

Definition 1. (Summary Graph) Let $G = (V, E)$ be a graph and $\Phi = \{V_1, V_2, V_3, \dots, V_k\}$ be a partition of G such that $\bigcup_{i=1}^{k} V_i = V$ and $\forall i \neq j$: $V_i \cap V_j = \emptyset$. The summary of G based on $\Phi$ is $G_s = (V_s, E_s)$ where $V_s = \Phi$ and $E_s = \{(V_i, V_j) | \exists\, u \in V_i \wedge \exists\, v \in V_j \wedge (u, v) \in E\}$.

Fig. 1 shows a graph and its structural summary. As shown in Fig. 1(a), vertices a, b and c of the original graph were grouped together and made a super-node (blue one) in the summary graph (Fig. 1(b)). The summary graph has four super-nodes corresponding to four dashed ovals of the original graph along with four super-edges. For more clarity, super-nodes have the same color as their corresponding groups in the original graph.

Table 1. Symbols and abbreviations which are used in this text

| *Notation* | *Interpretation* |
|---|---|
| $G$ | Graph |
| $G_S$ | Summary graph |
| $c_i$ | Importance of $i^{th}$ attribute |
| $V_i$ | $i^{th}$ super-node |
| $\alpha$ | Contribution of the structure in the resulting summary |
| Den | Density |
| $G_{S_{den}}$ | The density of summary graph $G_S$ |
| $G_{S_{ent}}$ | The entropy of summary graph $G_S$ |
| $p_{ijn}$ | The percentage of vertices in super-node $V_j$ that have value $a_{in}$ on attribute $a_i$ |
| $\text{ent}(a_i, V_j)$ | The entropy of super-node $V_j$ on attribute $a_i$ |
| $sim(v_i, v_j)$ | Similarity of two vertices $v_i$ and $v_j$ |
| $sim_{st}(v_i, v_j)$ | Structural similarity of two vertices $v_i$ and $v_j$ |
| $sim_{si}(v_i, v_j)$ | Attribute-based similarity of two vertices $v_i$ and $v_j$ |
| $sim_{si}(v_i, v_j, h)$ | Similarity of two vertices $v_i$ and $v_j$based on attribute $a_h$ |
| $val(v_i, a_k)$ | The value of single-valued attribute $a_k$ on vertex $v_i$ |
| $vals(v_i, a_k)$ | The values of multi-valued attribute $a_k$ on vertex $v_i$ |

### 3.3 Attribute-Based Summarization

To demonstrate this kind of summarization, it is necessary to define attributed graphs. The definition of an attributed graph according to [29] is as follows:

Definition 2. (Attributed Graph) An attributed graph is defined as 4-tuple $G = (V, E, \Sigma, F)$ where $V = \{v_1, v_2, \dots, v_n\}$ is a set of n nodes, $E = \{(v_i, v_j) | 1 \leq i, j \leq n \text{ and } i \neq j\}$ is a set of m edges, $\Sigma = \{a_1, a_2, \dots, a_k\}$ is a set of k attributes. Attributes of node $v_i \in V$ is denoted by $[a_1(v_i), a_2(v_i), \dots, a_k(v_i)]$ where $a_j(v_i)$ is the observation value of $v_i$ on attribute $a_j$. The set $F = \{f_1, f_2, \dots, f_k\}$ denotes a set of k functions and each $f_i: V \mapsto \text{dom}(a_i)$ assigns each node $v_j \in V$ an attribute value in the domain $\text{dom}(a_i)$ of attribute $a_i$ $(1 \leq i \leq k)$.

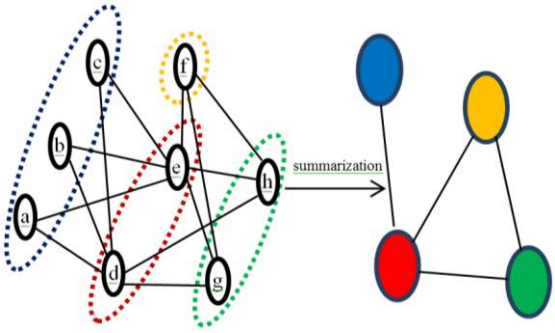


Fig. 1. (a) Original graph (left) (b) Summary graph (right)

The definition of an attribute-based summary is as follows:

Definition 3. (Attribute-based Summary) For a given graph $G = (V, E)$ let:

- Every node has an attribute set $A_v = \{a_1, a_2, \dots, a_k\}$.
- $\Phi = \{V_1, V_2, \dots, V_m\}$ is a partition on V.
- The user is interested in attributes $A_u = \{a_{i_1}, a_{i_2}, \dots, a_{i_j}\}$ where $A_u \subseteq A_v$.
- All vertices inside $V_i$ have the same value for each attribute of $A_u$.

Then $G_S = (V_S, E_S)$ where $V_S = \Phi$ and $E_S = \{(V_i, V_j) | \exists u \in V_i \text{ and } \exists v \in V_j \text{ s.t. } (u, v) \in E\}$ is an attributed-based summary.

Fig. 2(a) displays an attributed graph and one of its augmented graphs is shown in Fig 2(b). In some attributed-based summarization methods of a given graph, a new graph called augmented graph is constructed. In this new constructed graph, some virtual edges are added due to the attribute-based similarity of vertices. For example, the graph shown in Fig. 2(b) is an augmented graph of the one depicted in Fig. 2(a). The weight of an edge in the augmented graph is summation of structural and attribute-based similarities of its two end vertices, but they are not necessarily equal contributions. The structural and attribute-based summaries of this graph is shown in Figs 3(a) and 3(b).

### 3.4 Hybrid Summarization

The definition of hybrid summarization (summarization based on both structure and vertex attributes) is as follows:

Definition 4. (Hybrid Summary) For a given graph $G = (V, E)$, if:

1. Every node has an attribute set $A_v = \{a_1, a_2, \dots, a_k\}$.
2. $\Phi = \{V_1, V_2, \dots, V_k\}$ is a partition on V.
3. The user is interested in attributes $A_u = \{a_{i_1}, a_{i_2}, \dots, a_{i_j}\}$ where $A_u \subseteq A_v$.

Then $G_S = (V_s, E_s)$ will be a hybrid summary provided that the following conditions are met:

1. $G_S$ is a structural summary as previously mentioned.
2. All vertices inside $V_i$ have equal value for every attribute in $A_u$.
3. The edge density of super-nodes is higher than a given threshold.
4. The edge density between super-nodes is lower than a given threshold.

The hybrid summary of the graph shown in Fig. 2(a) is demonstrated in Fig 3(c). The summary is generated based on both structural and attribute-based similarities. The hybrid summary as shown in Fig 3(c), is different from the other two summaries.

In the following section, the proposed method and its components are described in details.

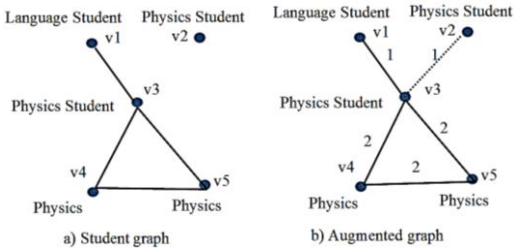


Fig. 2. Student graph and one of its augmented graphs

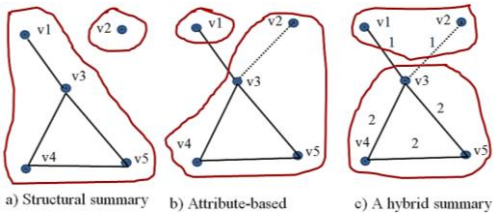


Fig. 3. Three different summaries of the student graph

## 4. The Proposed Method

The paradigm of the new method is shown in Fig 4. This paradigm consists of six components (four procedures, one system and criterion): 1-preprocessing 2-partitioning 3-knowledge-based reasoning 4- more summarization feasibility check 5- preparing for further summarization and 6- stopping criterion. These components have been illustrated in more details below.

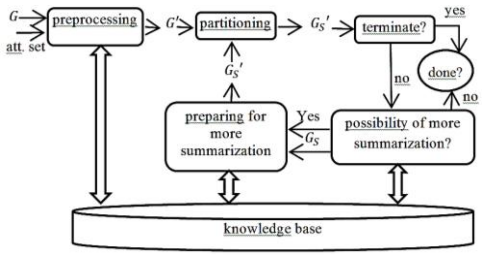


Fig. 4. The graph summarization paradigm

The new proposed method is a general one that covers three of the above-mentioned summarization. In fact, all three different kinds of summarization, only differs in terms of similarity measure, which is used merely in the partitioning component. The proposed paradigm has been summarized in Algorithm 1.

**Algorithm 1:** Proposed summarization method

**Input:** Graph $G$, attribute set A, importance of attributes C, summary size k, structure contribution $\alpha$.
**Output:** summary graph $G_s$.

```
1  begin
2  Construct the ontology;
3  Preprocess G;
4  Partition G into super-nodes and super-edges;
5   While(Gs's size <>k or the summary
          quality is not good)
6     if(more summarization is possible)
       Change the summary for further
7        summarization;
8      Resummarize the summary;
9  end.
```

### 4.1 Preprocessing and Constructing the New Augmented Graph

In the preprocessing procedure, the graph and the user goal are received as the input. The user goal is expressed based on vertex attributes. The degree of an attribute's relevance to the user goal is determined by the user or by communication with the knowledge base. Unrelated attributes are removed and relevance degrees of attributes are calculated for future applications. Based on the given graph, a new graph is constructed.

The values of categorical fields and their categories or intervals are determined. Noisy values are cleansed and vertices whose errors are beyond a given threshold are removed. New attributes, which can be defined based on current attributes, are introduced and their values are calculated and stored in the knowledge base. The defined attributes represent new concepts and are useful to curtail the summary. The topics of this subsection have been summarized in Algorithm 2.

### 4.2 Knowledge-Based Reasoning

The knowledge base in this paradigm contains all information about nodes and their attributes, which is required for summarization. The knowledge covers concepts, individuals and their relationships and attributes values, related to nodes. In fact, the ontology of the domain is maintained in the knowledge base.

An attributed graph can contain several ontologies. For example, for business trips, there are two ontologies called geographic and financial ontologies. The geographic ontology contains information about the location of sources and destinations of trips while financial ontology describes prices, different currencies and payment methods. Most components of the proposed paradigm are engaged with the knowledge base.

**Algorithm 2:** Preprocessing

**Input:** graph $G$, attribute set A, importance of attributes C; summary graph $G_s$.
**Output:** the preprocessed graph

```
begin
 Remove attributes that are less related to the summarization goal;
 Remove nodes whose errors are greater than a given threshold;
 Add new required attributes;
 Categorize the domain values of numerical attributes;
 Determine the order of components for hierarchal attributes;
 Unify words by considering synonyms;
end.
```

The knowledge can be represented by semantic networks, rules and first-order logic. The first and second ones have shortcomings compared to the last one [30]. The first order logic is not suitable for this purpose due to its un-decidability. A special type of the first order logic called Descriptive logic is suitable and could be used to represent the knowledge [30]. This representation supports reasoning that obtains implicit knowledge from the explicitly stored knowledge. That is, two nodes that are dissimilar in terms of the explicit knowledge may be similar with respect to the implicit knowledge.

## 4.3 Partitioning

A graph is partitioned into smaller parts based on a specific similarity criterion. Vertices are grouped together based on criteria such as structural similarity, attribute-based similarity or both. In all of these three partitioning cases, the similarity of two vertices is calculated and then vertices are partitioned into smaller parts based on the similarity. In fact, similar vertices are categorized in one group and dissimilar vertices in different groups.

## 4.4 Stopping Criterion

Summarization process should be terminated based on a criterion. Criteria such as size and quality of summary can be used for stopping summarization process. In some cases, summarization may be stopped since further summarization is impossible. When the summary size is decided by the user, stopping criterion is obvious, obtaining a summary of that size. Otherwise, stopping criterion can be defined based on the summary quality. As the quality increases, summarization is continued. The quality of summary, depending on the type of summary, could be defined in terms of density, entropy or a combination of both as follows:

- **Structural summary:** The quality of this kind of summary is measured in terms of density. The definition of density for a summary graph with m super-nodes is as follows:

$$\text{density}(\{V_i\}_{i=1}^{m}) = \sum_{i=1}^{m} \frac{|\{(v_p,v_q)|\, v_p,v_q \in V_i \text{ and } (v_p,v_q)\in E\}|}{|E|} \quad (1)$$

- **Attribute-based summary**: In this kind of summary, the entropy measure is used to evaluate the quality of summary. The definition of entropy for a summary graph with m super-nodes and k associated vertex attributes is as follows:

$$\text{entropy}(\{V_i\}_{i=1}^{m}) = \sum_{i=1}^{k} \frac{w_i}{\sum_{p=1}^{t} w_p} \sum_{j=1}^{m} \frac{|V_j|}{|V|} \text{entropy}(a_i, V_j) \quad (2)$$

where

$$\text{entropy}(a_i, V_j) = -\sum_{n=1}^{n_i} p_{ijn} \log_2 p_{ijn}$$

and $p_{ijn}$ is the percentage of vertices in the super-node $V_j$ with value $a_{in}$ on attribute $a_i$.

- **Hybrid summary**: The quality of a hybrid summary is measured in terms of density and entropy. In fact, Formula (3) is used for this purpose.

$$\text{Qual}(G_S) = \alpha * G_{S_{den}} / (1-\alpha) G_{S_{ent}} \quad (3)$$

Where $\alpha$ and $(1-\alpha)$ are contributions of structure and vertex attributes in the quality of the summary, respectively. The value of $\alpha$ is determined based on the importance of the structure in graph summarization.

Obviously, a good summary is the one with dense super-nodes and few interconnections. Thus, the quality of a hybrid summary is directly related to density and indirectly associated with entropy. The importance of the structure and attributes in summarization are not necessarily the same. For this reason, we multiplied density and entropy by $\alpha$ and $(1-\alpha)$, respectively.

## 4.5 More Summarization Feasibility Check

Sometimes users are interested in a summary of a specific size. Thus, the summarization process should be continued to obtain a summary of that size. However, it is not possible to summarize a graph to obtain a summary of an expected size. Hence, a criterion is necessary to check the possibility of further summarization.

For further summarize, there are a number of options such as combining intervals (ranges) of attribute values, promotion in a hierarchical attribute (e.g. student by human) or replacing a group of attributes by a newly introduced attribute (concept). The substitution of sub-types by super-types in a hierarchical attribute can be continued to reach the highest level of the hierarchy structure. Another criterion for stopping summarization process is based on the summary quality measures such as density and entropy, which are presented by Equations (1) and (2). In fact, when the summary size is not specified by the user, summarization process is stopped when the summary quality is not increased. The summary quality is defined based on a compromise between density and entropy. Algorithm 3 is used for further summarization feasibility check.

## 4.6 Preparing for more Summarization

The degree of summarization depends on the attribute set. By changing the attribute set or attributes, the summary size changes. Changing attributes such as promotion in a hierarchical field or combining adjacent intervals of values affects the summary. Thus,

hierarchical and categorical fields provide the best fields for changing the level of summarization. Replacing some attributes by a new attribute also increase the level of summarization. Algorithm 4 is proposed for this purpose.

## 5. Experimental Results

In this section, tools, datasets and computations of the proposed method and finally the results are proposed.

### 5.1 Gephi

We used Gephi to extract structural information and visualization of the graph. Gephi is the leading visualization and exploration software for all kinds of graphs and networks.

**Algorithm 3:** More summarization feasibility check

```
Input:  Summary graph G_s;
Output: A boolean value;
        begin
1       if( (at least one of the hierarchical
2           attributes is not at the highest
        level)
                          or
           (combining at least two adjacent
                  interval values is possible)
                          or
             (introducing a new attribute is
                        possible)
3         return true;
4       else return false;
5       end.
```

Gephi is an open-source and free software and its latest version (0.9.1) for Windows was used in this study. Gephi can import data to social networks also Facebook or Twitter and generate a graph and clusters.

**Algorithm 3:** Preparing for further summarization

```
Input:  A Summary graph;
Output: A Summary graph;
1   begin
2    let a be the less important attribute;
3    if(a is a hierarchical attribute)
       consider a higher component of this
         field;
4    else if( a is a numerical attribute)
5        decrease the number of its intervals;
6    else if(introducing a new attribute is
               possible)
7    introducing a new attribute  and add it;
8   end.
```

### 5.2 Dataset

We generated a graph with 1000 nodes and 2500 edges using R-Mat method and associated five attributes of age, gender, country, level of education and spoken languages to its vertices. These attributes were assigned values based on existing statistics for social networks. With the aim of obtaining graph structure information such as the number of connected components and their sizes, we developed a program for this purpose. The graph contained 185 sub-graphs of the sizes 813, 2, 2, 2 and 1. It is needless to say that the last one has the occurrence of 181. We visualized this graph using Gephi, as shown in Fig 5.

The structural features of this graph were extracted by Gephi was shown in Table 2.

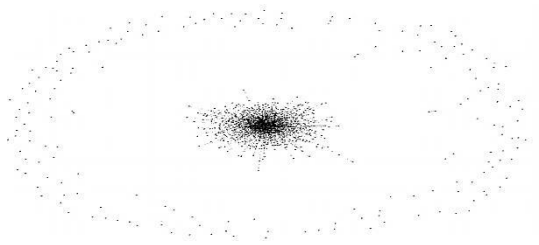

Fig. 5. Visualization of the graph by Gephi

Table 2. The extracted structural features of the graph using Gephi

| average degree | diameter | Density | connected components | modularity |
|---|---|---|---|---|
| 4.94 | 14 | 0.005 | 185 | 0.371 |

### 5.3 Computations

This section describes how to calculate the similarity of a pair of nodes, a node and a super-node or two super-nodes. That is, the necessary calculations for the hybrid summarization are presented here. The similarity of two vertices based on the structure and attributes is calculated as follows:

$$\mathrm{sim}(v_i, v_j) = \alpha \times \mathrm{sim}_{st}(v_i, v_j) + (1-\alpha) \times \mathrm{sim}_{si}(v_i, v_j) \quad (4)$$

In Equation (4), $\mathrm{sim}_{st}$ and $\mathrm{sim}_{si}$ are the structural and attribute-based similarities, respectively. These two functions are calculated as follows:

$$\mathrm{sim}_{st}(v_i, v_j) = \begin{cases} 0 & w[i][j] = 0 \\ 1 & w[i][j] = 1 \end{cases} \quad (5)$$

where $w$ is the adjacency matrix of the given graph. If every node has $k$ attributes, then the attributed-based similarity is calculated as follows:

$$\mathrm{sim}_{si}(v_i, v_j) = \sum_{l=1}^{k} c_l \times \mathrm{sim}_{si}(v_i, v_j, a_l) \quad (6)$$

$$\text{s.t. } 0 \le c_l \le 1 \text{ and } \sum_{l=1}^{k} c_l = 1$$

where $c_l$ is the importance of $l^{th}$ attribute, which it is provided by the user and $\mathrm{sim}_{si}(v_i, v_j, a_l)$ is the similarity of two vertices based on an attribute $a_l$ which is computed as follows:

$$\mathrm{sim}_{si}(v_i, v_j, a_l) = \begin{cases} 0 & a_l \text{is single\_valued} \wedge \mathrm{val}(v_i, a_l) \neq \mathrm{val}(v_j, a_l) \\ 1 & a_l \text{is single\_valued} \wedge \mathrm{val}(v_i, a_l) = \mathrm{val}(v_j, a_l) \\ \frac{|\mathrm{vals}(v_i,a_l) \cap \mathrm{vals}(v_j,a_l)|}{|\mathrm{vals}(v_i,a_l) \cup \mathrm{vals}(v_j,a_l)|} & a_l \text{ is a multi\_valued attribute} \end{cases} \quad (7)$$

where $\mathrm{val}(v_i, a_l)$ is the value of attribute $a_l$ on vertex $v_i$ and $\mathrm{vals}(v_i, a_l)$ is a set of values for attribute $a_l$ on vertex $v_i$. The attribute-based similarity of two vertices in terms of a multi-valued attribute is calculated based on Jaccard similarity as depicted in Equation (7).

The similarity of a node and a super-node is computed using Equation (8).

$$\mathrm{sim}(V_p, v_l) = \alpha \times \mathrm{sim}_{st}(V_p, v_l) + (1-\alpha) \times \mathrm{sim}_{si}(V_p, v_l) \quad (8)$$

The structural similarity of a super-node and a node is the number of edges between the node and the super-node divided by the super-node size. Thus, Equation (9) can be utilized for this purpose.

$$\mathrm{sim}_{st}(V_p, v_l) = \frac{|\{u|u \in V_p \text{ and } (u,v_l) \in E\}|}{|V_p|} \quad (9)$$

The attribute-based similarity of a super-node and a node indicates the summation of attribute-based similarity of the vertex and the super-node on associated attributes. Thus, Equation (10) can be used for this purpose.

$$\mathrm{sim}_{si}(V_p, v_l) = \sum_{i=1}^{k} c_i \mathrm{sim}_{si}(V_p, v_l, a_i) \quad (10)$$

Where

Table 3. Summaries with 5 super-nodes and different values of α

| $G_S$ summary | 916, 47, 18, 14, 5 | 821, 92,54, 25, 8 | 558, 426, 9, 5, 24 | 549, 396, 50, 3, 2 | 415, 247, 244, 79, 15 |
|---|---|---|---|---|---|
| $\alpha$ | 1.0 | 0.75 | 0.5 | 0.25 | 0.0 |
| **density** | 0.4 | 0.107 | 0.230 | 0.172 | 0.123 |
| **entropy** | 353.947 | 297.09 | 278.950 | 257.926 | 199.914 |
| $Qual(G_S)$ | 0.001 | 0.0003 | 0.0008 | 0.0006 | 0.0006 |

$$\mathrm{sim}_{si}(V_p, v_l, a_i) = \frac{|\{u|u \in V_p \text{ and } val(u,a_i) = val(v_l,a_i)\}|}{|V_p|} \quad (11)$$

The similarity of two super-nodes is calculated as follows:

$$\mathrm{sim}(V_p, V_q) = \frac{1}{|V_q|} \sum_{i=1}^{|V_q|} (\mathrm{sim}(V_p, v) | v \in V_q) \quad (12)$$

The quality of a hybrid summary is measured by $(\alpha * density)/((1-\alpha) * entropy)$ and it can be used as a stopping criterion in the summarization algorithm.

## 5.4 Implementation

The proposed method was implemented in Java for evaluation. We developed this program by designing classes such as Graph, SummaryGraph, Node, Edge, SuperNode and SuperEdge to construct and summarize a graph. The SummaryClass has a number of methods to summarize a graph and calculate the density and entropy of the generated summary.

## 5.5 Time Complexity

In the proposed method, the dominant time belongs to the **partitioning** component. The preprocessing and construction of knowledge base are performed once and it is also obvious that their run-times is less than the run-time of the **partitioning** component. The knowledge is stored in a tree structure. The run-time of components such as **terminate**, **possibility of further summarization** and **preparing for further summarization** are also less than the run-time of **partitioning** component. The run-time of **partitioning** is $O(n^2)$ and since this component is repeated a maximum of $n$ times, the time complexity of the proposed method will be $O(n^3)$, where $n$ is the number of vertices in the graph.

## 5.6 Results

We generated a number of summaries using the proposed method, as depicted in Tables 3, 4, and 5, to demonstrate the efficiency of the proposed method, as we did in our previous works [31] –[32]. The first row of these tables shows the size of super-nodes. For example, in Table 3, the second column of the first row indicates that the summary has five super-nodes of sizes 916, 47, 18, 14 and 5. Other rows represent $\alpha$, density, entropy and the quality of each summary, respectively. As shown in Tables 3, 4 and 5, the values of density, entropy and the quality of each summary are calculated for different values of $\alpha$.

To assess the quality of summaries based on the contribution of the structure in the summary, we changed the value of α from 0 to 1 with an incremental rise of 0.25 in each step. The quality of the summary based on α parameter is presented in Figures 6, 7 and 8.

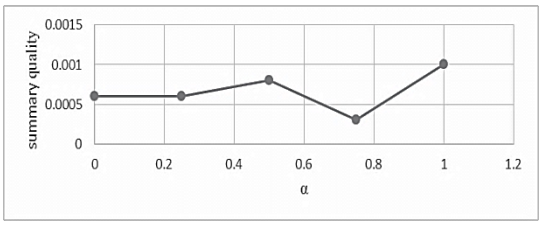


Fig. 6. The quality of summary graph in terms α parameter.

Table 4. Summaries with 10 super-nodes and different values of α

| $G_S$ summary | 835, 60,40, 20,19, 13, 5, 3 | 818,8, 1, 3, 5, 3, 6, 3 | 529, 436, 15, 5, 5, 4, 4, 2 | 523,36, 39, 33, 5, 15, 7, 3 | 395,29, 181,4, 45, 3, 9, 3 |
|---|---|---|---|---|---|
| $\alpha$ | 1.0 | 0.75 | 0.5 | 0.25 | 0.0 |
| density | 0.25 | 0.208 | 0.137 | 0.109 | 0.063 |
| entropy | 315.306 | 297.71 | 269.148 | 241.748 | 182.515 |
| $Qual(G_S)$ | 0.007 | 0.007 | 0.005 | 0.004 | 0.003 |

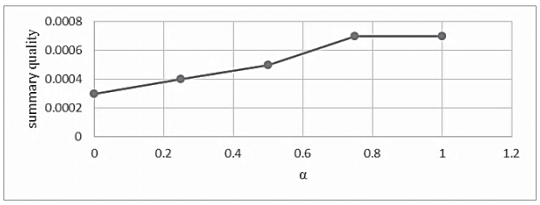


Fig. 7. The quality of summary graph in terms of α parameter.

Table 5. Summaries with 10 super-nodes and different values of α

| $G_S$ summary | 830, 47, 32, 31, 31, 8, 7, 5, 5, 4 | 864, 68, 16, 12, 10, 8, 8, 6, 6, 2 | 491, 378, 33, 32, 22, 19, 14, 6, 3, 2 | 289, 244, 236, 134, 38, 18, 16, 15, 7, 3 | 472, 275, 80, 60, 52, 18, 15, 15, 10, 3 |
|---|---|---|---|---|---|
| $\alpha$ | 1.0 | 0.75 | 0.5 | 0.25 | 0.0 |
| density | 0.2 | 0.177 | 0.100 | 0.049 | 0.068 |
| entropy | 315.220 | 322.050 | 232.163 | 159.422 | 196.894 |
| $Qual(G_S)$ | 0.006 | 0.005 | 0.004 | 0.003 | 0.003 |

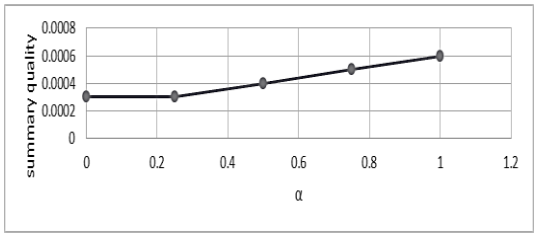


Fig. 8. The quality of summary graph in terms of α parameter.

## 6. Discussions

In order to assess the quality of summaries, we generated summaries of sizes 5, 8 and 10 and changed the value of α from 0 to 1 with a 0.5 increase in each step. These summaries and their features are presented in Tables 3, 4 and 5. The quality of summaries in terms of $\alpha$ are presented in Figures 6, 7 and 8.

Table 3 shows the summaries of size 5 with different values of α. The values of density, entropy and the newly proposed criterion, Qual( $G_S$ ), are calculated for each summary according to every value of α.

Figures 6, 7 and 8 show the quality of summaries in terms of the value of α.

As can be seen, by increasing the value of α, the value of entropy rises , but this increase is not significant and does not affect the quality of summary graph.

In these figures, by increasing the value of α the quality of summary is also improved. Hence, the quality of the summary graph is enhanced by increasing the contribution of structure in the similarity of vertices. The results suggest that the relationship of vertices in this graph is based on the connection of vertices to their similarity. In this way, we can change the value of α to generate a summary with the highest quality.

According to Figures 6, 7 and 8 and also the proposed criterion for summary quality, the best summary of the constructed graph has anα value of 1. This is in agreement with the features of the graph discovered by the Gephi tool and the program developed in Java. The best value of α, which corresponds to a summary of the high quality, can be learned by the algorithm. This is a topic to be pursued in future works.

## 7. Conclusions and Future Works

In this paper, we proposed a new method for summarizing a graph in an interactive and knowledge-based manner. The proposed method could summarize a graph based on users' needs. The proposed method was able to summarize a graph based on the structure, attributes or both. In this regard, ontology was used for graph summarization as it generated a summary of the right size based on user's needs. The user can determine the contributions of structure and vertex attributes in generating the summary. We proposed a criterion to measure the quality of a hybrid summary. The proposed criterion allows comparing the quality of summaries.

With the aim of evaluating the proposed method, we generated summaries of different sizes and values of α for a synthetic graph. The values of density, entropy and the proposed quality criterion were calculated for each generated summaries. To extract and visualizing the structural information of the graph, we also used the Gephi tool and an application developed in Java.

The experimental results showed that the proposed method generated a hybrid summary of higher quality. The experimental results were consistent with the graph topological structure, obtained from the above-mentioned tools.

We plan to extend the proposed method to summarize graph streams based on sliding windows to enable the monitoring of a graph stream. Using this method, hybrid summaries are compared to each other syntactically and semantically. A further research venue would be summarizing multiple graph streams.

## References


[1] U. Kang, "Mining Tera-Scale Graphs: Theory, Engineering and Discoveries," 2012.

[2] "Facebook active users." [Online]. Available: https://www.yahoo.com/news/number-active-users-facebook-over-230449748.html.

[3] S. Navlakha, R. Rastogi, and N. Shrivastava, "Graph summarization with bounded error," Proc. 2008 ACM SIGMOD Int. Conf. Manag. data - SIGMOD '08, p. 419, 2008.

[4] K. LeFevre and E. Terzi, "GraSS: Graph Structure Summarization," Proc. 2010 SIAM Int. Conf. Data Min., pp. 454–465, 2010.

[5] N. Zhang, Y. Tian, and J. M. Patel, "Discovery-driven graph summarization," 2010 IEEE 26th Int. Conf. Data Eng. (ICDE 2010), pp. 880–891, 2010.

[6] M. A. Beg, M. Ahmad, A. Zaman, and I. Khan, "Scalable approximation algorithm for graph summarization," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10939 LNAI, pp. 502–514, 2018.

[7] Y. Liu, T. Safavi, N. Shah, and D. Koutra, "Reducing large graphs to small supergraphs: a unified approach," Soc. Netw. Anal. Min., vol. 8, no. 1, 2018.

[8] Y. Tian, R. A. Hankins, and J. M. Patel, "Efficient Aggregation for Graph Summarization," pp. 567–579.

[9] Y. Bei, Z. Lin, and D. Chen, "Summarizing scale-free networks based on virtual and real links," Phys. A Stat. Mech. its Appl., vol. 444, no. 2, pp. 360–372, 2016.

[10] H. Cheng, Y. Zhou, and J. X. Yu, "Clustering Large Attributed Graphs: A Balance between Structural and Attribute Similarities," ACM Trans. Knowl. Discov. Data, vol. 5, no. 2, pp. 1–33, 2011.

[11] M. Riondato, D. García-Soriano, and F. Bonchi, "Graph summarization with quality guarantees," Data Min. Knowl. Discov., vol. 31, no. 2, pp. 314–349, 2017.

[12] X. Liu, Y. Tian, Q. He, W.-C. Lee, and J. McPherson, "Distributed Graph Summarization," Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manag. - CIKM '14, pp. 799–808, 2014.

[13] C. Chen, C. X. Lin, M. Fredrikson, M. Christodorescu, X. Yan, and J. Han, "Mining graph patterns efficiently via randomized summaries," Proc. VLDB Endow., vol. 2, no. 1, pp. 742–753, 2009.

[14] S. Hosseini, H. Yin, M. Zhang, Y. Elovici, and X. Zhou, "Mining Subgraphs From Propagation Networks Through Temporal Dynamic Analysis," in 2018 19th IEEE International Conference on Mobile Data Management (MDM). IEEE, 2018.

[15] P. Pourashraf, N. Tomuro, S. B. Shouraki, "From Windows to Logos: Analyzing Outdoor Images to Aid Flyer

Classification", International Conference Image Analysis and Recognition, Springer, Cham, pp. 175-184, 2018.
[16] U. Von Luxburg, "A Tutorial on Spectral Clustering," Stat. Comput., vol. 17, no. March, pp. 395–416, 2007.
[17] I. Dhillon, Y. Guan, and B. Kulis, "A unified view of kernel k-means, spectral clustering and graph cuts", Technical Report, Computer Science Department, University of Texas at Austin, pp. 1–20, 2004.
[18] B. Auffarth, "Spectral Graph Clustering," Univ. Barcelona course Rep. Tech. Av. Aprendizaj Univ. Politec. Catalunya, pp. 1–12, 2007.
[19] S. Uw, A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," Adv. Neural Inf. Process. Syst. 14, pp. 849–856, 2002.
[20] D. Zhou and C. J. C. Burges, "Spectral clustering and transductive learning with multiple views," Proc. 24th Int. Conf. Mach. Learn. - ICML '07, pp. 1159–1166, 2007.
[21] S. Smyth and S. White, "A spectral clustering approach to finding communities in graphs," Proc. 5th SIAM Int. Conf. Data Min., pp. 76–84, 2005.
[22] J. Liu, C. Wang, M. Danilevsky, and J. Han, "Large-scale spectral clustering on graphs," IJCAI Int. Jt. Conf. Artif. Intell., pp. 1486–1492, 2013.
[23] C.-D. Wang, J.-H. Lai, and P. S. Yu, "Dynamic Community Detection in Weighted Graph Streams," Proc. 2013 SIAM Int. Conf. Data Min., pp. 151–161, 2013.
[24] G. T. Prabavathi and V. Thiagarasu, "Overlapping Community Detection Algorithms in Dynamic Networks: An Overview," Int. J. Emerg. Technol. Comput. Appl. Sci., pp. 299–303, 2013.
[25] P. Ben Sheldon, "Community Detection Algorithms: a comparative evaluation on artificial and real-world networks D . Phil student report," Other, pp. 1–27, 2010.
[26] W. Wang and W. N. Street, "A novel algorithm for community detection and influence ranking in social networks," ASONAM 2014 - Proc. 2014 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., no. Asonam, pp. 555–560, 2014.
[27] O. Benyahia et al., "Community detection in dynamic graphs with missing edges To cite this version: HAL Id: hal-01590597 Community Detection in Dynamic Graphs with Missing Edges," 2017.
[28] C. Chen, X. Yan, F. Zhu, J. Han, and P. S. Yu, "Graph OLAP - Towards Online Analytical Processing on Graphs," Data Mining, 2008. ICDM'08. Eighth IEEE Int. Conf., pp. 103–112, 2008.
[29] Y. Wu, Z. Zhong, W. Xiong, and N. Jing, "Graph summarization for attributed graphs," Proc. - 2014 Int. Conf. Inf. Sci. Electron. Electr. Eng. ISEEE 2014, vol. 1, pp. 503–507, 2014.
[30] S. Grimm, P. Hitzler, and A. Abecker, "Knowledge Representation and Ontologies Logic, Ontologies and Semantic Web Languages," Semant. Web Serv., pp. 51–105, 2007.
[31] N. Ashrafi Payaman, M. R. Kangavari, "GSSC: Graph Summarization based on both Structure and Concepts", International Journal of Information & Communication Technology Research, vol. 9, no. 1, pp. 33-44, 2017.
[32] N. Ashrafi Payaman, M. R. Kangavari. "Graph Hybrid Summarization", Journal of AI and Data Mining, vol. 6, no. 2 pp. 335-340, 2018.

**Nosratali Ashrafi Payaman** received his B.Sc. degree in software engineering from Kharazmi University in 1999 and his M.Sc. degree in computer science from Sharif University of Technology in 2002. He is currently a Ph.D. candidate in software engineering at Iran University of Science and Technology and is also a faculty member of Kharazmi University. His current main research interests include analysis and design of algorithms, graph summarization and software vulnerability.

**Mohammad Reza Kangavari** received his B.Sc. degree in mathematics and computer science form Sharif University of Technology (1982), his M.Sc. degree in computer science from Salford University (1989), and his Ph.D. degree in computer science from the University of Manchester (1994). He is currently an Associate Professor at the Department of Computer Engineering, Iran University of Science and Technology. His research interests include intelligent systems, machine learning, and wireless sensor networks.

# Modeling the Inter-arrival Time of Packets in Network Traffic and Anomaly Detection Using the Zipf's Law

Ali Naghash Asadi
Trustworthy Computing Laboratory, School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
aliasadi@comp.iust.ac.ir
Mohammad Abdollahi Azgomi*
Trustworthy Computing Laboratory, School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
azgomi@iust.ac.ir

Received: 15/Jan/2018 Revised: 10/Aug/2018 Accepted: 06/Oct/2018

**Abstract**

In this paper, a new method based on the Zipf's law for modeling the features of the network traffic is proposed. The Zipf's law is an empirical law that provides the relationship between the frequency and rank of each category in the data set. Some data sets may follow from the Zipf's law, but we show that each data set can be converted to the data set following from the Zipf's law by changing the definition of categories. We use this law to model the inter-arrival time of packets in the normal network traffic and then we show that this model can be used to simulate the inter-arrival time of packets. The advantage of this law is that it can provide high similarity using less information. Furthermore, the Zipf's law can model different features of the network traffic that may not follow from the mathematical distributions. The simple approach of this law can provide accuracy and lower limitations in comparison to existing methods. The Zipf's law can be also used as a criterion for anomaly detection. For this purpose, the TCP_Flood and UDP_Flood attacks are added to the inter-arrival time of packets and they are detected with high detection rate. We show that the Zipf's law can create an accurate model of the feature to classify the feature values and obtain the rank of its categories, and this model can be used to simulate the feature values and detect anomalies. The evaluation results of the proposed method on MAWI and NUST traffic collections are presented in this paper.

**Keywords:** Network Traffic Modeling; Inter-arrival Time; Anomaly Detection; DoS Attack; The Zipf's Law.

## 1. Introduction

Today, network traffic analysis and examination of its different aspects have great importance. Researchers need the artificial traffic to evaluate the efficiency and security of networks. In other words, they require a traffic generator to simulate the actual traffic. For this purpose, researchers must study the actual traffic without anomaly and extract patterns from it; then they can generate traffics that follow from the extracted patterns, and import them into the network. By doing this, they can monitor the networks based on efficiency and security. Furthermore, researchers can create abnormal traffic by identifying the normal traffic behavior and examine the network security with it. The models can also be used to provide security in the network. One way to detect anomalies is that we identify the normal behavior of network traffic and define any deviation from it as an anomaly. In other words, any deviation from the model obtained from normal traffic is considered as an anomaly. Thus the modeling of different features of the network traffic is important in anomaly detection.

In this paper, the Zipf's law is proposed to model the features of the network traffic and detect anomalies. This law can model the network traffic using less information from it, and provide accurate simulation using the resulting model. Furthermore, we show that the model obtained from the Zipf's law can detect anomalies. To do this, firstly we create a normal traffic collection and model its features using the Zipf's law. The resulting normal model can be used to simulate the normal network traffic, and to detect anomalies by examining the deviations from the normal traffic. The main contribution of this paper can be summarized in four major categories:

1) Modeling the features of the network traffic using less information.

2) The Zipf's law can model different features of the network traffic that may not follow from mathematical distributions.

3) Using the model obtained from the Zipf's law, we can simulate the features of the network traffic and detect anomalies which affect these features.

4) The Zipf's law is very suitable for detecting attacks that have frequent sequential patterns, for example, Denial-of-Service (DoS) attack.

The rest of this paper is organized as follows. In section 2, the Zipf's law, the reasons for network traffic analysis, and anomaly detection approaches will be introduced. In section 3, the related works will be presented in the field of network traffic modeling and anomaly detection. In section 4, firstly we prove mathematically that all data sets can be converted to the data sets following from the Zipf's law. Then the proposed method based on the Zipf's law, for modeling network traffic and using it for simulation and

* Corresponding Author

anomaly detection, is presented. In this section, the traffic collections and experiment method will also be defined. Furthermore, we will show how the number of categories must be determined in the Zipf's law. In section 5, the experiments related to the traffic modeling and anomaly detection are evaluated more accurately. Furthermore, the Zipf's law is compared with Benford's law and the entropy theory. In the final section, the results are evaluated.

## 2. Background

### 2.1 The Zipf's law

George Kingsley Zipf, professor of linguistics at Harvard University, in 1949 reached conclusions about the words and their frequencies in the text by studying the words in books. His initial result showed that if you count all words in a book and sort them in descending order, you will see that the rank of each word is inversely proportional to its frequency. In other words, the number of each word appeared in the text is inversely proportional to its rank. This relationship is known as the Zipf's law. According to this law, a word with the rank of 1 appears twice more than a word with the rank of 2 in the text. It also appears three times more than a word with the rank of 3 [1].

This law shows the relationship between a frequency *F* and a rank *R* (Eq. (1)) [1][2]. Based on this relation, the frequency of each word multiplied by its rank will be almost a constant value in the text. In this relation, *N* is the total number of the words and *A* is the constant value close to 0.1 *(0<A<1)*. Figure 1 shows the frequency of each rank in the Zipf's law [1][2].

$$\frac{(R_i \times F_i)}{N} = A \quad (1)$$

It can also be defined as a logarithmic relation (Eq. (2)) [1][2]. The logarithmic form of the Eq. (1) and its graph are important and they will be used in this paper.

$$\log R_i + \log \frac{F_i}{N} = \log A \quad (2)$$

This law can also be used in other topics. It is very strange how and why a simple relation occurs in many complex topics. For example, the relationship between the Zipf's law and the coherence property of the urban system (e.g. the city size distribution) has been studied [3][4]. Furthermore, the relationship between this law and the distribution of user-generated passwords has been investigated [5]. Researchers, with the concrete knowledge of password distributions, suggest a new metric based on the Zipf's law for measuring the strength of password datasets. In [6], the capacity scaling law of a device-to-device (D2D) caching network according to the Zipf popularity distribution has been studied. Also, in [7], researcher investigate the use of Benford's law and the Zipf's law to distinguish between humans using keystroke biometric systems and non-humans for auditing application. The law is also used for the simulation of a number of hits on the World Wide Web, page rank prediction, and viral email detection [8][9].

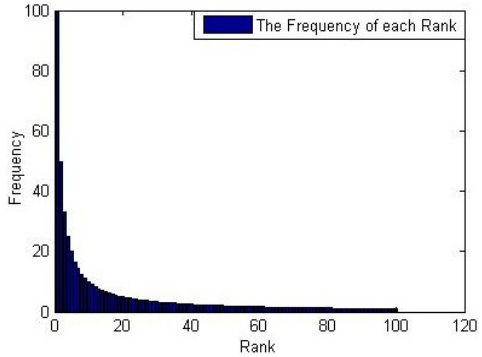


Fig. 1. Zipf's law diagram

### 2.2 Network Traffic Analysis

Researchers are interested to display and understand the network traffic. Network traffic analysis includes the processes of capturing and evaluating the network traffic [10]. This concept is also known as network analysis, protocol analysis, and packet sniffing.

Network traffic analysis is done for different objectives and it can provide important information on the user behavior patterns. Network managers can get a better understanding of the networks with these patterns. An important purpose of the network traffic analysis is the modeling of different features of the network traffic. The models created from different features of the network traffic can be used in various fields, including development of the normal and attack traffic simulators, detection of the anomalies and attacks, study of the service performance, examination of the network security policies, projection of the future network requirements, and prevention of the network traffic monitoring.

Development of the normal and attack traffic simulators is an important issue in network traffic modeling. Due to the lack of proper traffic collection for experiments, researchers are trying to create traffic generators with the most similarity to the actual traffic. These traffics must be generated based on the needs of the network managers; therefore, the global traffics that are available cannot be suitable. Traffic generators can have an important role in the evaluation of the service performance of a newly established network, and the monitoring of the security policies before implementing them. Researchers have created many network traffic generators with different characteristics, including Harpoon [11], D-ITG [12], TCPReplay [13], and Avalanche [14]. Network traffic analysis and modeling can provide adequate information about network requirements in the future, and network administrators can define proper policies and programs for these requirements.

The attackers are interested to access network traffic. They receive critical information in this way. To prevent it, some systems add artificial traffic to their actual traffic [15]. For this purpose, the artificial traffic should be added to the actual traffic to conceal important information and the actual behavior of the network. Furthermore, the artificial traffic should not allow that attackers distinguish it from the actual traffic. Therefore,

network traffic modeling is important for generating the artificial traffics that are similar to the actual traffic. In [16], the problems of those who want to simulate internet traffic have been described; however, the solutions have also been proposed to address the problems [17].

The network traffic modeling can also be used to detect anomalies and attacks. For example, by modeling the different features of the normal network traffic and comparing it with actual traffic, any deviation from the normal model can be considered as an anomaly [18]. Therefore, accuracy in modeling for accurately identifying anomalies and attacks and reducing false alarm is very important. Traffic analysis is used to detect errors in the network and also can help to understand the main reason of the error and its effect on communication between users.

### 2.3 Anomaly and Attack Detection

Intrusion detection systems (IDS) can be classified into three approaches: signature-based, behavior-based, and a hybrid approach [19]. A signature detection system identifies patterns of traffic to detect a malicious activity, while an anomaly detection system compares the activities occurring with normal activities [19]. The hybrid approach uses a combination of approaches. The main advantage of a signature detection system is that the known attacks can be detected with a low false alarm rate. These systems require a signature for every attack. The main advantage of an anomaly detection system is that the unknown attacks can be detected, but these systems have the high percentage of false alarms.

Network traffic anomalies can be classified into outages, flash crowds, attacks, and measurement failures [20]. Network outages include any network failure event or temporary misconfigurations. Flash crowds are mainly caused by the sudden increase in the users of a special service. Network attacks can typically be any kind of intentional failures, including flood-based denial-of-service events. Measurement failures can be caused by problems with the data collection infrastructure itself.

A general approach for anomaly detection is to define a normal behavior and observe the actual behavior of the system and compare them. Any deviations from normal behavior should be considered as an anomaly [18]; however, there are many challenges to this simple approach. The key point is the difficulty of defining a normal behavior that should include any possible normal behaviors. Also, anomalies and methods of attacks change rapidly. Another major challenge is the massive amounts of data. Anomaly detection techniques require efficient computing to handle the large incoming data. In addition, data are usually online, therefore they need online analysis. Another important issue that arises because of the high volume input is that even a low percentage of false alarms can make analysis overwhelming.

## 3. Related Work

### 3.1 Network Traffic Modeling

Researchers have introduced many models on network traffic. According to their goals, they have investigated various features of the network traffic and modeled some of them. For example, important features of the Telnet, SMTP, HTTP, and FTP protocols are modeled in [21]. Before it, researchers had described most features of the network traffic with the Poisson distribution, but the others showed that except for user-initiated TCP session arrivals, other TCP connection arrivals don't follow from the Poisson distribution [21].

In [22], the inter-arrival time of the TCP and UDP packets are studied by the Kolmogorov-Smirnov method to match with the mathematical distributions. In this paper, the Pareto and Weibull distributions are introduced as the most appropriate distributions for representing the inter-arrival time of the TCP and UDP packets.

In a network, packets can be transferred with a specific maximum size which is determined by the MTU parameter, and thus large packets should be divided into small ones. It leads to the inaccurate simulation of the actual packet size. In [23], researchers have been studied the packet size to estimate the probability density function that has a minimum difference with the packet size distribution graph.

Before considering the concept of self-similarity, the Poisson was the most important distribution for modeling the network traffic. The concept means that the statistical graphs of features will never change at different scales. Researchers show that the features of the network traffic have the self-similarity feature, and so they cannot be modeled by the Poisson distribution [24][25]. In [26], methods have been developed for modeling of self-similar traffic and loading process of telecommunication networks. After discovering this, the Weibull became the most important distribution to describe the different features of the network traffic [27]. In addition, the majority of features of the network traffic are long-tail, and it leads to more importance of the Weibull distribution because it can show the long-tail behavior for some shape and scale parameters. In [28][29], the role of the Weibull distribution in internet traffic modeling has been explained. In [29], researchers empirically show that despite the variety of data networks in size, number of users, applications, and load, the inter-arrival times of normal flows correspond to the Weibull distribution.

In [30], the network jitter has been modeled by mathematical distributions. Jitter can affect the quality of service. So jitter can be modeled to identify the factors that cause it. In these papers, the packet jitter has been studied with considering its path nodes. For this purpose, type and number of nodes are used for modeling the packet jitter.

### 3.2 Anomaly and Attack Detection in Network Traffic

Fernandes et al. have been proposed a survey to review the most important aspects related to anomaly detection [31]. They show that the anomaly detection methods can be divided into six categories, including statistical methods, clustering methods, finite state machine methods, classification-based methods, information theory, and evolutionary computation.

As mentioned, intrusion detection systems (IDS) are divided into three categories: signature-based, behavior-based, and hybrid approach [19]. Snort and Bro are signature-based intrusion detection systems which have rules to detect attacks [32]. SPADE, NIDES, PHAD, and ALAD are examples of behavior-based intrusion detection systems which are based on the statistical models and produce the anomaly detection alarms when a large deviation from normal behavior occurs [32].

Behavior-based intrusion detection systems, to evaluate the deviation from normal behavior, need to define the fast and accurate criteria that be able to work with a large volume of data. For this purpose, many methods have been proposed, including Benford's law and entropy theory. For example, in [33], Benford's law is used as a criterion to detect anomalies in the inter-arrival-time of SYN packets (the TCP flow initiator) which follows from a Weibull distribution with the shape parameter less than one. The difference between the inter-arrival-time of the actual flows and the Weibull distribution can be used to detect anomalies affecting the flow of SYN packets. This paper showed that the random variable of the Weibull distribution follows the Benford's law, so easily and without loss of generality, the Weibull conformance test can be replaced with the first-digit test which is less complicated. Also in [34], this law has been used to detect anomalies in the UDP packets and flows.

The entropy theory is more used in the anomaly detection process [35][36][37][38][39][40]. For example, in [35], the entropy theory is used to detect anomalies created by the SYN and Port_Scan attacks. In other words, the actual network traffic is compared with the basic distribution defined for normal traffic with the help of the entropy theory. In [36], important features of packets are selected for defining rules that prevent from the DoS attacks with the help of the entropy theory. In [37], the performance of IDS based on data mining and K-means algorithm is modified by using entropy theory. In our paper, the behavior-based method based on the Zipf's law will be used to detect anomalies.

## 4. The Zipf's Law in Modeling and Anomaly Detection

In section 3, different models of the features of the network traffic have been introduced. In these models, the researchers compare the features of the network traffic with mathematical models and distributions, and if they match, those models will be used as the best representation of the normal behavior of features. However, this approach has limitations. For example, the features that don't follow from a specific mathematical distribution can't be modeled. Furthermore, the model obtained from a traffic collection may not be usable in other traffic collections. In this paper, the Zipf's law is used to solve the above problems. This law can provide a normal behavior model from feature values with ranking the different categories.

In this section, firstly we prove that all data sets can be converted to the data sets following from the Zipf's law. Then we use this law for modeling the network traffic and using it for simulation and anomaly detection. Furthermore, we will show how the number of categories must be determined in the Zipf's law.

### 4.1 Following all Data Sets from the Zipf's Law

According to the Zipf's law (Eq. (1)), the rank of each category is inversely proportional to its frequency. In other words, the frequency of each category ($f_i$) can be obtained by multiplying the sum of all frequencies ($N$) in the constant value ($A$), and then divided it by its rank ($r_i$). According to this law, the frequency of a category with the rank of 1 ($f_1$), Maximum frequency) is equal to multiplication the sum of all frequencies in the constant value ($f_1 = N \times A$). Furthermore, this value ($f_1$) can be calculated by multiplying the rank in the frequency other categories ($f_1 = r_i \times f_i$).

In this subsection, we prove that all data sets (even those that don't follow from the Zipf's law) can be converted to the data sets following from the Zipf's law by changing the definition of categories. For this purpose, we assume that there is a finite data set $DS$, and the categories $C = \{c_1, c_2, \dots, c_i, \dots, c_n\}$ has been defined for it. We show the frequency of each category with $f(c_i)$. Furthermore, according to Eq. (3), the categories with the rank of $k_{th}$ and their frequency are shown with $r_k$ and $f(r_k)$, respectively. If the $C$ can be converted to the $C' = \{c'_1, c'_2, \dots, c'_j, \dots, c'_m\}$ that $f(r_1) = k \times f(c'_j)$, it can be proven that the $DS$ follows from the Zipf's law. The $r_k$ includes the categories ($c'_j$) which have the rank of $k_{th}$, and their frequency is equal to the maximum frequency ($f(r_1)$), the rank of 1) division by the $k$. We assume that the $r_k$ can be null except $k = 1$.

$$r_k = \{c'_j | f(r_k) = f(c'_j) = {f(r_1)}/{k}\} \quad (3)$$

Now, with the above assumptions, we prove our hypothesis. We assume that the categories defined from the $DS$ have the same frequency ($f(c_i) = K, K$ is constant). In this case, our hypothesis can be proved by aggregating categories. The $C = \{c_1, c_2, \dots, c_n\}$ can be converted to the $C' = \{c_1, c_{2,3}, c_{4,5,6,7}, \dots\}$that the $r_k (k = 1,2,4,8, \dots)$ may have only one member, the others are null, and $f(r_1) = k \times f(r_k)$. The $r_1$ is equal to the aggregation of the maximum number of the $c_i s$. It is important that we can always define the $C = \{c_1, c_2, \dots, c_n\}$ with $f(c_i) = 1 (K = 1)$. So all data sets can follow from the Zipf's law. Figure 2 shows an example for this purpose.

We show that the rank of each category is the best description of that category, and by having it and the total number of samples ($N$) (in this paper, packets) can calculate

the frequency of each category according to the Zipf's law (Eq. (1)). In this paper, the rank of each category is our model, and it can be used in the simulation and anomaly detection.

## 4.2 The Proposed Method

In this subsection, the proposed methods for modeling, simulation, and anomaly detection are defined. The summary of our methods is shown in Figure 3. To perform the experiments, firstly we create a data set of the selected feature values. Then we model the data set with the Zipf's law. This model can be used in the simulation and anomaly detection. As mentioned in subsection 4.1, all data sets can follow from Zipf's law with changing the definition of their categories. In the Zipf's law, the number and range of categories must be specified. This part is the most important step in the Zipf's law. If the number and range of categories were properly selected, an accurate model would obtain that can simulate features and detect anomalies with high efficiency. In subsection 4-3, we explain how to determine the number of categories. In the next step, according to Eq. (4), the maximum value of the data set is subtracted from the minimum value and then it is divided by the number of categories. By doing so, the range of categories is determined.

$$Range = \frac{Max(Feature\_Values) - min(Feature\_Values)}{The\ Number\ of\ Categories} \qquad (4)$$

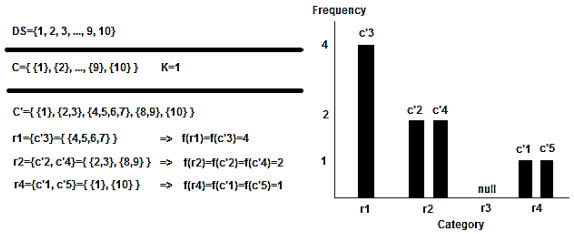


Fig. 2. An example of changing the definition of categories for following a data set from the Zipf's law

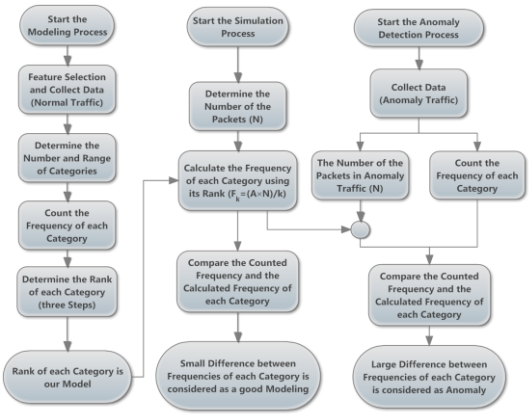


Fig. 3. our proposed methods for modeling, simulation and anomaly detection

Then the frequency of each category must be counted by viewing the data set. After obtaining the frequency of each category, the rank of each category should be determined. As mentioned, the rank of categories is used as a model in the Zipf's law. In order to determine the rank of categories, three steps have been defined:

- Step 1: The frequency of categories is sorted in ascending order, then the first category is assigned the rank of 1, and the last category is assigned a rank equal to the number of categories. Categories that have a frequency of zero or one will have the rank of zero.
- Step 2: The rank of categories that have the same frequency must be set to the lower rank. For example, if the fourth and fifth categories, which have the rank of 4 and 5 respectively, have the same frequency, their rank will be 5.
- Step 3: As mentioned in the definition of the Zipf's law, the frequency of a category with the rank of 1 will be twice more than the frequency of a category with the rank of 2, and three times more than the frequency of a category with the rank of 3. To comply with this law, the pseudocode of Algorithm 1 is used. In this pseudocode, the rank of each category, obtained from step 2, will be modified. For this purpose, if the frequency of the category with the rank of k is smaller than the frequency of the first category divided by $k$ and is greater than the frequency of the first category divided by $k+1$, the rank of the category is suitable for it. Otherwise, the $k$ value will increase, and step 3 will be done again. In this pseudocode, the $NC$ is the number of categories. The $R(c)$ is the rank of the $c_{th}$ category. The $FCnt(c)$ is the counted frequency of the $c_{th}$ category.

```
algorithm step3 of ranking in the Zipf's law is
   input: the number of categories (NC),
                the counted frequencies (FCnt(c)),
          the rank of each category from step2 (R(c))
   output: the rank of each category in step3 (R(c))
start
   for c := 2 to NC
            k := R(c)
            flag := 'true'
            while (flag == 'true')
                if ((FCnt(c) - FCnt(1)/k) equal or near-equal to zero)
                        R(c) := k
                        flag := 'false'
                else_if ((FCnt(c)<FCnt(1)/k) and (FCnt(c)>FCnt(1)/(k+1)))
                        R(c) := k + 1
                        flag := 'false'
                k := k + 1
            end_while
   end_for
end
```

Algorithm. 1 Step 3 of the ranking

```
algorithm calculating the frequency of each category in the Zipf's
law is
   input: the number of categories (NC),
                the number of packets (TP),
          the rank of each category (R(c))
   output: the frequency of each category (FCal(c))

start
   sum := 0
   remain := TP - sum
   while (remain > 0)
                for c := 1 to NC
                   if (R(c) == 0)
                           FCal(c) := 0
                   else
                           FCal(c) := FCal(c) + (A * remain/R(c))
                           sum := sum + FCal(c)
                end_for
                remain := TP – sum
   end_while
end
```

Algorithm. 2 Calculating the frequency of each category

By doing the above steps, a model obtains from the normal behavior of the network traffic feature. We show that this model is accurate. To do this, the frequency of each category is calculated by using the Zipf's law and the obtained ranks, and their results are compared with actual normal counted frequencies. In other words, the calculated frequencies are compared with the counted frequencies. As mentioned, the counted frequencies obtain from actual normal traffic, and the calculated frequencies obtain using the pseudocode of Algorithm 2. To perform the comparison, the frequency graph is used in both simple and logarithmic types. The SSE difference criterion is also used to show the difference between the counted and calculated frequencies. This criterion is calculated by using Eq. (5).

$$SSE = \sum_{i=1}^{NC} (log_{10} FCal(r_i) - log_{10} FCnt(r_i))^2 \qquad (5)$$

In this relationship, the *NC* is the number of categories. Also, *FCal* and *FCnt* are the calculated and counted frequencies respectively. The $r_i$ is the rank of the $i_{th}$ category. To prove that the Zipf's law can provide an accurate model, the *SSE* value of each category must be small.

The pseudocode of Algorithm 2 is used to calculate the frequency of each category. This pseudocode will be executed until the total of the calculated frequencies is equal to the number of the required packets. In order to evaluate the model, the total of the calculated frequencies is considered equal to the total number of packets of the actual normal network traffic, but this value in anomaly detection is considered equal to the total number of packets of the abnormal network traffic. Furthermore, the total of the calculated frequencies in the simulation is considered equal to the number of the required packets. As can be seen in the pseudocode of Algorithm 2, only the rank of each category is used for calculating the frequency of each category. In this pseudocode, the frequency of each category is calculated by dividing the number of the required packets on its rank. The constant value of *A* can be set to *0.1*. however, for greater precision in the *A* value, according to Eq. (6), the counted frequency of each category can be multiplied by its rank and divided by the total number of packets, and then their sum can be divided by the number of the non-zero categories. This value can also be considered as the A value that will always be a number close to *0.1*. In this relation, the *NCNZ* is the number of the non-zero categories.

$$A = \frac{1}{N \times NCNZ} \times \sum_{i=1}^{NC} (FCnt(i) \times r_i) \qquad (6)$$

We use from the pseudocode of Algorithm 2 to show the accuracy of the model, the network traffic simulation, and anomaly detection. We can use the calculated frequencies for network traffic simulation. In other words, we specify the number of required packets for simulating and then use from the resulting model and the pseudocode of Algorithm 2 for calculating the frequency of each category and generate packets according to it. In this paper, packet generation isn't done actually. We just calculate the frequency of each category using the model and show that these frequencies are similar to frequencies of the actual normal traffic. In other words, we know how many packets must be in each category in the total number of packets in the normal network traffic.

For detecting anomalies, the same above activities are done. For example, suppose the feature and number of categories have been selected and the rank of each category has been extracted from normal traffic. In other words, we have the normal behavior model. Now we will examine abnormal traffic with this model. For this purpose, the frequency of each category must be counted in abnormal traffic. Then according to the total number of abnormal traffic packets, the expected frequency of each category in normal traffic is calculated by the pseudocode of Algorithm 2 and the normal model. Then the calculated frequency of each category from the normal model is compared with the counted frequency of each category from abnormal traffic. Categories which their frequency shows a large difference according to the SSE value are considered as the anomaly. If the counted frequency of each category is not proportional to its rank, the SSE value will be increased.

We add the TCP_Flood and UDP_Flood attacks to the normal traffic of MAWI and NUST to create two abnormal traffic collections. There are these attacks in the NUST

traffic collection separately. Since we know the location of attacks in the traffic collections, the deviation in the areas proves that the proposed method works properly. The SSE criterion is used to detect anomalies. For this purpose in Eq. (5), the calculated frequency from model normal is compared with the counted frequency from abnormal traffic.

## 4.3 Determining the Number of Categories

Determining the exact number of categories is the most important step in the Zipf's law. If it is properly selected, an exact model will be achieved that leads to high efficiency in the anomaly detection. The number of categories influences the speed of processes and the accuracy of results. To determine the number of categories, a balance must be established between the increase in the number of categories and the decrease in the number of categories with zero frequency (meaningless categories).

The increase in the number of categories leads to the lower speed of the modeling, simulation and anomaly detection processes. It also increases categories with zero frequency. For example, the range of 0.5 byte is meaningless for packet size because it doesn't get a decimal value. Moreover, if the maximum and minimum values of the data set are too far away, and the values aren't uniformly distributed over the entire range, many categories with zero frequency will be produced which may not have a role in the modeling, simulation and anomaly detection. Nevertheless, the increase in the number of categories will increase the accuracy of results. For example, assume the frequency of packets that their packet size is between 80 and 100 bytes is 1000. If we divide this range into two parts, the frequency of packets that their packet size is between 80 and 90 bytes may be 800. This example clearly shows the effect of increasing the number of categories in modeling. Moreover, if an anomaly event occurs in packets that their packet size is 95 bytes, anomaly detection in the range of between 90 and 100 bytes will be done much easier than the range of between 80 and 100 bytes. Nevertheless, the increase in the number of categories or the decrease in the range of categories is preferred because the modeling process is done only once.

The minimum range of categories can vary in different features. For example, the minimum range of categories in packet size feature is one byte because it cannot be a decimal value. However, the minimum range of categories in inter-arrival time feature is different and it can be selected based on the accuracy and unit of time (e.g., seconds or milliseconds). We can also use several ranges for each experiment. Figure 1 shows that the frequency of early categories is much more than the frequency of other categories. We can choose a smaller range for these categories and a larger range for others; however, we use the same range for all categories of each experiment in this paper. This range is calculated by Eq. (4).

## 4.4 Experimental Results

### 4.4.1 Traffic Collections

NUST Traffic collection [41] is used in our paper. It has been collected at the National University of Sciences and Technology in Pakistan, and its benign and attack packets have been marked. There is header and payload information of more than 6 million TCP packets and 1 million UDP packets in this traffic collection. Also, there are three subsets (home, isp and soho) in it that show traffic capture sources. Furthermore, there are attack packets include TCP_Flood, TCP_Port_Scan, and UDP_Flood in this traffic collection. These attack packets can be injected into normal packets for evaluation of intrusion detection systems.

The other traffic collection that is used in our paper is MAWI [42]. MAWI, the measurement and analysis on the WIDE internet, captures the packets from a trans-pacific backbone connecting a combination of general and academic hosts and servers to the internet. The MAWI traffic collection is a public traffic extracted from the MAWI working group traffic archive. Each file of the traffic collection, collected over a 150 Mbps trans-pacific backbone link, consists of the first 96 bytes of all IP packets sent over the link from 14:00 to 14:15 every day. We just used one minute of the collection collected in 2014/11/01. There is header and payload information of more than 3 million TCP packets and 3 thousand UDP packets in it. We have removed attacks from this traffic collection using the Snort software. Therefore, it can be used as a normal traffic collection.

### 4.4.2 The Zipf's Law in Modeling and Simulation

In this part, we show two graphs for each experiment (Figure 4). The number of categories is 10000 in all experiments. The first graph compares the counted frequency and the calculated frequency of each category. The vertical axis is the frequency of each category, and the horizontal axis is the number of each category. The second graph compares the logarithm of the counted frequency and the calculated frequency of each category. The vertical axis is the logarithm of the frequency of each category, and the horizontal axis is the logarithm of the rank of each category. In the second graph, the logarithm of the frequency of each category has been shown in order of its rank.

Experiments have been conducted on the inter-arrival time of packets. Details of these experiments have been shown in Table 1. The inter-arrival time of packets obtains through the subtraction of arrival time of each packet from the arrival time of the previous packet. Because these values are very small, all values are multiplied by the large constant value. The results of experiments have been shown in Figure 4.

As can be seen in Table 1, the SSE value for all experiments is very small. Also in Figure 4's graphs, the frequency of each category, counted from the normal traffic, conforms to the frequency of the same category, calculated from the normal model. Therefore, this model can be used in network traffic simulation. In other words, the number of packets of each category in a normal

artificial traffic can be specified by this model. Also, we know how many packets with a certain feature value must be generated in network traffic simulation. This simulation provides the highest accuracy with minimum information.

### 4.4.3 The Zipf's Law in Anomaly Detection

In this part, we show two graphs for each experiment (Figure 5). The number of categories is 10000 in all experiments. The first graph compares the calculated SSE of each packet in normal traffic. The second graph compares the calculated SSE of each packet in abnormal traffic. The calculated SSE value of each category is assigned to all packets of it. So if a category is abnormal, all packets of it will be abnormal. In both graphs, the vertical axis is the SSE value of each packet, and the horizontal axis is the number of each packet.

We create abnormal traffic collections by adding 3000 packets of the TCP_Flood and UDP_Flood attacks to normal traffic collections. These packets are added from the $200{,}000_{th}$ packet onwards of the TCP and UDP normal traffic collections. The packets of these attacks are sent at certain time intervals, 10 packets per second for 300 seconds, to achieve their objectives (disrupting the target system). For this reason, a specific inter-arrival time in the traffic collection will significantly increase. Thus, according to the Zipf's law, one or more categories will have better rank, and the changing rank against the normal rank is easily detected by the Zipf's law. In other words, the SSE value of these categories increases significantly. As can be seen in the second graph, the SSE value from the $200{,}000_{th}$ packet onwards increased significantly. However, as mentioned, some of the normal packets are known as an anomaly.

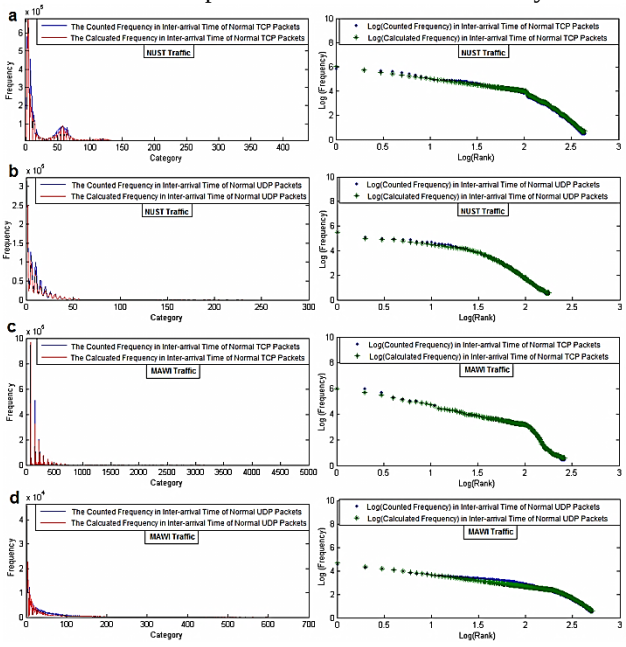


Fig. 4. The counted and calculated frequencies graphs in both simple and logarithmic in different traffic collections: (a) TCP packets of NUST collection; (b) UDP packets of NUST collection; (c) TCP packets of MAWI collection; (d) UDP packets of MAWI collection;

Table 1. Details of the modeling experiments

| Traffic Collection | Inter-arrival time | Number of packets | Minimum value | Maximum value | Number of categories with a non-zero frequency | Range of categories | A | SSE |
|---|---|---|---|---|---|---|---|---|
| NUST | TCP packets | 6276243 | zero | 7935 | 438 | 0.7935 | 0.1293 | 0.0026 |
| | UDP packets | 1472908 | 0.001 | 98.141 | 177 | 0.0098 | 0.2131 | 0.0024 |
| MAWI | TCP packets | 3768306 | zero | 416 | 524 | 0.0416 | 0.2569 | 0.0028 |
| | UDP packets | 327740 | zero | 3543 | 831 | 0.3543 | 0.1387 | 0.0233 |

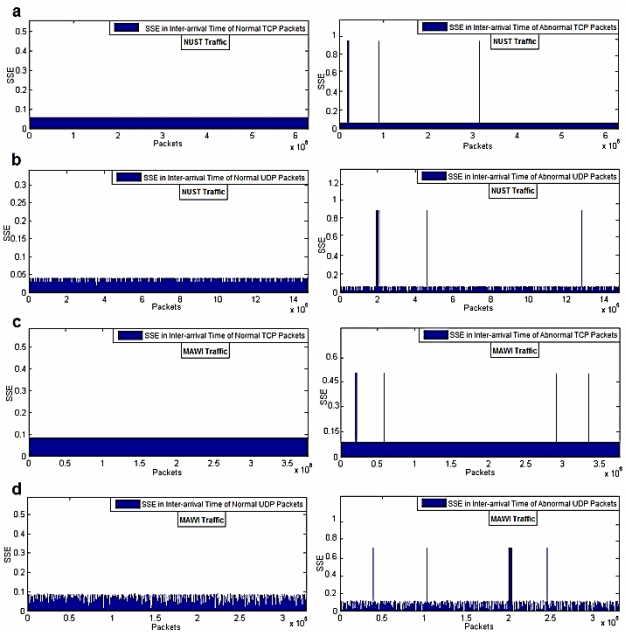


Fig. 5. The calculated SSE graphs in both normal and abnormal traffic collections in different traffic collections: (a) TCP packets of NUST collection; (b) UDP packets of NUST collection; (c) TCP packets of MAWI collection; (d) UDP packets of MAWI collection;

## 5. Evaluation

### 5.1 Evaluation of Experiments in the Different Number of Categories

In this paper, the Zipf's law was used for modeling, simulation and anomaly detection in network traffic. At the end of this paper, the proposed method is evaluated more accurately. The evaluations were done on a system with a quad-core CPU and 4 GB RAM.

#### 5.1.1 Evaluation of the Modeling and Simulation

The modeling time and SSE Criteria are used to evaluate the modeling and simulation. In other words, the elapsed time for modeling and the SSE value which shows the difference between the calculated frequency from the model and the counted frequency from actual traffic, are used to evaluate them. According to the algorithms 1 and 2, the modeling time depends on the number of packets and categories. It is obvious that the lower values for both

criteria indicate the better efficiency of the experiment. Summary of results has been shown in Table 2. As mentioned, the number and range of categories have an important role in modeling and simulation in the Zipf's law. We show the importance of this issue by repeating the experiments in the different number of categories, including 1000, 5000, and 10,000 categories.

As can be seen in Table 2, by increasing the number of categories, the modeling time increases; however, the SSE value doesn't change much with this increase. The modeling time and the SSE value are appropriate in a fewer number of categories, but increasing the number of categories can provide more information from events occurring in the network traffic. For example, assume the frequency of packets that their packet size is between 80 and 100 bytes is 1000. If we divide this range into two parts, the frequency of packets that their packet size is between 80 and 90 bytes may be 800. This example clearly shows the effect of increasing the number of categories in modeling. However, increasing the number of categories may lead to produce the categories with zero frequency which do not have a role in the modeling, simulation and anomaly detection. Therefore, the number of categories must be determined according to the needs. Table 3 shows the number of categories with a non-zero frequency in different number of categories.

As can be seen in Table 2, the SSE value is very low in all results; this means that the Zipf's law is effective in network traffic modeling and simulation. So by increasing the number of categories, the modeling and simulation are done accurately.

Table 2. The modeling and simulation experiments in the different number of categories

| Traffic Collection | Inter-arrival time | Number of categories | Range of categories | Modeling time (s) | SSE |
|---|---|---|---|---|---|
| NUST | TCP packets | 1000 | 7.9350 | 100.462 | 0.0052 |
| | | 5000 | 1.5870 | 400.259 | 0.0026 |
| | | 10000 | 0.7935 | 784.905 | 0.0026 |
| | UDP packets | 1000 | 0.0981 | 20.330 | 0.0042 |
| | | 5000 | 0.0196 | 92.727 | 0.0024 |
| | | 10000 | 0.0098 | 183.580 | 0.0024 |
| MAWI | TCP packets | 1000 | 0.4160 | 59.476 | 0.0029 |
| | | 5000 | 0.0832 | 245.049 | 0.0029 |
| | | 10000 | 0.0416 | 476.790 | 0.0028 |
| | UDP packets | 1000 | 3.5340 | 5.125 | 0.0748 |
| | | 5000 | 0.7086 | 21.698 | 0.0748 |
| | | 10000 | 0.3543 | 42.267 | 0.0233 |

Table 3. The number of categories with a non-zero frequency in the different number of categories

| Traffic Collection | Inter-arrival time | 1000 | 5000 | 10000 |
|---|---|---|---|---|
| NUST | TCP packets | 287 | 375 | 438 |
| | UDP packets | 53 | 92 | 177 |
| MAWI | TCP packets | 273 | 395 | 524 |
| | UDP packets | 532 | 667 | 831 |

### 5.1.2 Evaluation of the Anomaly Detection

The efficiency of anomaly and attack detection methods is often measured by two criteria: the detection rate and false alarm rate. To define the two criteria, two types of possible error which are their variables must be introduced. These two types of error are:

- False positive or false alarm error: A false positive error occurs when a normal event is detected as an attack event.
- False negative error: A false negative error occurs when an attack event is identified as a normal event.

$$False\ Positive\ Rate = \frac{Number\ of\ False\ Positive}{Total\ Number\ of\ Non\_Attacks} \quad (7)$$

$$False\ Negative\ Rate = \frac{Number\ of\ False\ Negative}{Total\ Number\ of\ Attacks} \quad (8)$$

$$Detection\ Rate = 1 - \frac{Number\ of\ False\ Negative}{Total\ Number\ of\ Attacks} \quad (9)$$

In this part, the two measures, the detection rate and false positive rate, are used to assess the proposed anomaly detection method. The elapsed time (detection time) to calculate the deviation is another criterion to assess this method. According to the algorithms 1 and 2, the detection time depends on the number of packets and categories. If the detection rate gets high value and the detection time and the false positive rate get low value, anomaly detection method will have better efficiency. Table 4 shows the results of the anomaly detection in the different number of categories. It is observed that by increasing the number of categories, the elapsed time to calculate the deviation increases, and also the detection rate and false positive rate are improved. For this reason, the number of categories must be determined according to the needs.

Table 4. The anomaly detection experiments in the different number of categories

| Traffic Collection | Inter-arrival time | Number of categories | Detection time (s) | FPR (%) | DR (%) |
|---|---|---|---|---|---|
| NUST | TCP packets | 1000 | 198.115 | 0.014 | 85.4 |
| | | 5000 | 525.944 | 0.006 | 95.27 |
| | | 10000 | 923.977 | 0.003 | 98.5 |
| | UDP packets | 1000 | 46.655 | 0.061 | 91.83 |
| | | 5000 | 121.930 | 0.040 | 96 |
| | | 10000 | 217.352 | 0.014 | 98.87 |
| MAWI | TCP packets | 1000 | 130.305 | 0.040 | 71.03 |
| | | 5000 | 315.321 | 0.013 | 92.97 |
| | | 10000 | 574.729 | 0.007 | 96.77 |
| | UDP packets | 1000 | 10.577 | 0.44 | 63.23 |
| | | 5000 | 27.821 | 0.19 | 93.23 |
| | | 10000 | 49.131 | 0.082 | 97.03 |

## 5.2 The Theoretical Comparison of the Zipf's Law with the Benford's Law and Entropy Theory

In section 3, we showed examples that used from Benford's law and entropy theory in anomaly detection. In this subsection, we propose more details of these laws and show advantages of the Zipf's law against them. Benford's law is an empirical law used in various fields. According to this law (Eq. (10)), the occurrence probability of the d digit as the first digit of any value in a data set isn't unexpectedly uniform. It has a logarithmic relationship [43].

$$P_d = \log_{10} \frac{d+1}{d} \quad (10)$$

We can say that Benford's law is a special case of the Zipf's law. Although Benford's law is a powerful law in anomaly detection, it has some limitations in this field. In this subsection, some limitations of Benford's law are presented and the Zipf's law is recommended due to fewer limitations and more benefits:

- Benford's law can only be done on numerical values because it needs to extract their first digit, but Zipf's law can be done on non-numeric values.
- Categories in Benford's law are the first digit of values, but they could be anything in the Zipf's law.
- The frequency graph of the first digit of feature values in Benford's law must be logarithmic, and this causes some data sets don't follow from Benford's law. The main advantage of the Zipf's law is that the frequency graph should not be necessarily logarithmic.
- Benford's law doesn't provide a model for feature values, and only when the first digit of feature values of network traffic is logarithmic, Benford's law can be used to detect anomalies. However, the Zipf's law can extract a model from feature values and use it in the simulation and anomaly detection.

In spite of all the advantages of the Zipf's law, Benford's law has advantages over the Zipf's law. The most important advantage of Benford's law is that it just counts the frequency of the first digit from the traffic collection and compares it with the logarithmic distribution, then it can detect anomalies. But for anomaly detection in the Zipf's law, the frequency of the normal model of each category is calculated and then compared with the counted frequency of the same category in the abnormal traffic. For this purpose, the categories must be defined and their normal rank extracted. So Zipf's law needs to extract the normal model.

Entropy is used in different science to study disorder and uncertainty. It emphasizes that normal systems have few disorders, and anomalies increase them; thus anomalies can be detected by this theory. In Eq. (11), $P(x_i)$ is the probability of the independent variable xi [44].

$$Entropy(X) = -\sum_i P(x_i) \times log_2 P(x_i) \quad (11)$$

The Zipf's law and the entropy theory are very similar to each other. Zipf's law extracts an order from a feature in the normal traffic collection and thereby can detect anomalies affecting it. The entropy theory has essentially been defined by order and disorder and can identify anomalies affecting a system.

In spite of the high similarity between the Zipf's law and the entropy theory, there are differences in the two approaches. For example, the entropy theory is more used in the anomaly detection process [35][36][37][38][39][40]. In other words, the entropy theory cannot use in the modeling and simulation processes. However, the Zipf's law can extract a model from normal traffic and use it in the simulation and anomaly detection. Thus the Zipf's law provides more capabilities than the entropy theory.

### 5.3 The Practical Comparison of the Zipf's Law with the Benford's Law

In the related works, we presented two papers that use the Benford's law to detect anomalies in the inter-arrival time of packets [33][34]. We select them for comparison because they use the same MAWI and NUST traffic collections, albeit with different law and assumptions. Table 5 shows the best results of the three papers for comparison. As can be seen, according to the descriptions of the subsection 5-2, the results of the three papers are almost the same. However, the detection time of our method is more than the other ones. This is because the Zipf's law needs to calculate the normal frequencies and to count the anomaly frequencies. But the Benford's law can only be used to detect anomalies, while the Zipf's law can be used as a perfect method for the modeling, simulation, and anomaly detection.

Table 5. Results obtained for comparing our proposed method based on the Zipf's law with the presented methods based on the Benford's law

| Traffic Collection | Inter-arrival time | Paper | Detection time (s) | FPR (%) | DR (%) |
|---|---|---|---|---|---|
| NUST | TCP packets | Our | 923.977 | 0.003 | 98.5 |
| | | [33] | 45.23 | 0.010 | 98.62 |
| | | [34] | - | - | - |
| | UDP packets | Our | 217.352 | 0.014 | 98.87 |
| | | [33] | - | - | - |
| | | [34] | - | - | - |
| MAWI | TCP packets | Our | 574.729 | 0.007 | 96.77 |
| | | [33] | 32.52 | 0.012 | 97.24 |
| | | [34] | - | - | - |
| | UDP packets | Our | 49.131 | 0.082 | 97.03 |
| | | [33] | - | - | - |
| | | [34] | 2.188 | 1.53 | 98.93 |

## 6. Conclusions

In this paper, the Zipf's law was used to model and simulate the normal behavior of network traffic and detect anomalies. The Zipf's law is an empirical law that has been used in various research topics. Some data sets may follow from the Zipf's law, but we proved that each data set can be converted to the data set following from the Zipf's law by changing the definition of categories. We used this law to model the inter-arrival time of TCP and UDP packets in the normal network traffic and then we proposed a method to detect anomalies by using the resulting model. For this purpose, the TCP_Flood and UDP_Flood attacks were added to the normal traffic collections and they were detected with high detection rate with the help of this law. We also showed that this model can be used to simulate the inter-arrival time of packets. Then we compared the Zipf's law with Benford's law and entropy theory in anomaly detection and showed that it can be used as a perfect method for the modeling, simulation, and anomaly detection.

For future works, we can examine other features of the network traffic which that may not follow from a particular mathematical distribution. The results of this work can be effective in detecting other attacks and anomalies. Furthermore, models created from other features can be used to develop a network traffic simulator.

## References


[1] G. Zipf, "Human behavior and the principle of least effort," The Economical Journal, vol. 60, no. 3, pp. 808-810, 1950.

[2] A. I. Saichev, Y. Malevergne and D. Sornette, Theory of Zipf's Law and Beyond, Springer-Verlag Berlin Heidelberg, 2010.

[3] S. Arshad, S. Hu and B. N. Ashraf, "Zipf's law and city size distribution: A survey of the literature and future research agenda," Statistical Mechanics and its Applications, vol. 492, no. 15, pp. 75-92, 2018.

[4] S. Arshad, S. Hu and B. N. Ashraf, "Zipf's law, the coherence of the urban system and city size distribution: Evidence from Pakistan," Physica A (2018), https://doi.org/10.1016/j.physa.2018.08.065.

[5] D. Wang, H. Cheng, P. Wang and G. Jian, "Zipf's Law in Passwords," IEEE Transactions on Information Forensics and Security, vol. 12, no. 11, pp. 2776-2791, 2017.

[6] A. Liu, V. Lau and G. Caire, "Capacity scaling of wireless device-to-device caching networks under the physical model," in IEEE International Symposium on Information Theory, Germany, 2017.

[7] A. Iorliam, A. T. Ho, N. Poh, S. Tirunagari and P. Bours, "Data forensic techniques using Benford's law and Zipf's law for keystroke dynamics," in International Workshop on Biometrics and Forensics, Norway, 2015.

[8] M. Jauhari, A. Saxena and J. Gautom, "Zipf's Law and Number of hits on the World Wide Web," Annals of Library and Information Studies, vol. 54, no. 2, pp. 81-84, 2007.

[9] L. Adamic and B. Huberman, "Zipf's Law and the Internet," in Glottometrics, 2007.

[10] B. R. Chang and H. F. Tsai, "Improving network traffic analysis by foreseeing data packet- flow with hybrid fuzzy-based model prediction," Expert Systems with Applications, vol. 36, no. 3, pp. 6960-6965, 2009.

[11] J. Sommers and P. Barford, "Self-Configuring Network Traffic Generation," in the 4th ACM SIGCOMM conference on Internet measurement, Italy, 2004.

[12] A. Botta, A. Dainotti and A. Pescape, "A tool for the generation of realistic network workload for emerging networking scenarios," Computer Networks, vol. 56, no. 1, pp. 3531-3547, 2012.

[13] "TCPReplay," [Online]. Available: http://tcpreplay.synfin.net/wiki. [Accessed 23 08 2018].

[14] "Network, devices & services testing-Spirent," [Online]. Available: http://www.spirent.com/. [Accessed 23 08 2018].

[15] W. M. Shbair, A. R. Bashandy and S. I. Shaheen, "A New Security Mechanism to Perform Traffic," in International Conference on Computational Science and Engineering, 2004.

[16] F. Sally and P. Vern, "Difficulties in simulating the internet," IEEE/ACM Transactions on Networking, vol. 9, no. 4, pp. 392-403, 2001.

[17] V. Paxon, "Strategies for sound internet measurement," in the 4th ACM SIGCOMM conference on Internet measurement, Italy, 2004.

[18] V. Chandola, A. Banerjee and V. Kumar, "Anomaly detection: a survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1-58, 2009.

[19] A. Patcha and J. M. Park, "An overview of anomaly detection techniques: existing solutions and latest technological trends," Computer Networks, vol. 51, no. 12, pp. 3448-3470, 2009.

[20] P. Barford, J. Kline, D. Plonka and A. Ron, "A signal analysis of network traffic anomalies," in the 2nd ACM SIGCOMM Workshop on Internet measurement, France, 2002.

[21] S. Luo and G. A. Marin, "Generating Realistic Network Traffic for Security Experiments," in IEEE SoutheastCon, USA, 2004.

[22] E. Garsva, N. Paulauskas, G. Grazulevicius and L. Gulbinovic, "Packet Inter-arrival Time Distribution in Academic Computer Network," ELEKTRONIKA IR ELEKTROTECHNIKA, vol. 20, no. 3, pp. 87-90, 2014.

[23] M. Fras, J. Mohorko and Z. Cucej, "Packet Size Process Modeling of Measured Self-similar Network Traffic with Defragmentation Method," in 15th International Conference on Systems, Signals and Image Processing, Slovakia, 2008.

[24] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, "the self-similar nature of Ethernet traffic (extended version)," IEEE/ACM Transactions on Networking, vol. 2, no. 1, pp. 1-15, 1994.

[25] X. An and L. Qu, "A Study Based on Self-Similar Network Traffic Model," in Sixth International Conference on Intelligent Systems Design and Engineering Applications, China, 2015.

[26] A. Pashko and V. Tretynyk, "Methods of the statistical simulation of the self-similar traffic," Advances in Intelligent Systems and Computing, vol. 754, no. 1, pp. 54-64, 2018.

[27] V. I. Strelkovskaya, T. I. Grygoryeva and I. N. Solovskaya, "Self-similar traffic in G/M/1 queue defined by the Weibull distribution," Radioelectronics and Communications Systems, vol. 61, no. 3, pp. 128-134, 2018.

[28] M. A. Arfeen, K. Pawlikowski, D. McNickle and A. Willig, "The Role of the Weibull Distribution in Internet Traffic Modeling," in 25th International Conference on Teletraffic Congress, China, 2013.

[29] L. Arshadi and A. H. Jahangir, "An empirical study on TCP flow interarrival time distribution for normal and anomalous traffic," International Journal of Communication Systems, vol. 30, no. 1, pp. 1-19, 2017.

[30] T. K. Bandhopadhya, M. Saxena and A. Tiwari, "Jitter's Alpha-Stable Distribution Behavior," Computer Technology and Electronics Engineering, vol. 3, no. 1, pp. 13-16, 2013.

[31] G. J. Fernandes, J. P. Rodrigues, L. F. Carvalho, J. F. Al-Muhtadi and M. J. Proença, "A comprehensive survey on network anomaly detection," Telecommunication Systems (2018), https://doi.org/10.1007/s11235-018-0475-8.

[32] "IDS Distribution," [Online]. Available: http://cs.fit.edu/~mmahoney/dist/. [Accessed 23 08 2018].

[33] L. Arshadi and A. H. Jahangir, "Benford's law behavior of Internet traffic," Journal of Network and Computer Applications, vol. 40, no. 1, pp. 194-205, 2014.

[34] A. N. Asadi, "An approach for detecting anomalies by assessing the inter-arrival time of UDP packets and flows using Benford's law," in 2nd International Conference on Knowledge-Based Engineering and Innovation, Tehran, 2015.

[35] Y. Gu, A. McCallum and D. Towsley, "Detecting Anomalies in Network Traffic Using Maximum Entropy Estimation," in the 5th ACM SIGCOMM conference on Internet measurement, USA, 2005.

[36] S. Honda, T. Nakashima and S. Oshima, "Entropy Based Analysis of Anomaly Access of IP Packets," in 3rd International Conference on Innovative Computing Information and Control, China, 2008.

[37] L. I. Han, "Research of K-MEANS Algorithm based on Information Entropy in Anomaly Detection," in Fourth International Conference on Multimedia Information Networking and Security, China, 2012.

[38] S. K. Gautam and H. Om, "Anomaly detection system using entropy based technique," in First International Conference on Next Generation Computing Technologies, India, 2015.
[39] A. A. Waskita, H. Suhartanto and L. T. Handoko, "A performance study of anomaly detection using entropy method," in International Conference on Computer, Control, Informatics and its Applications, Indonesia, 2016.
[40] D. Hong, D. Zhao and Y. Zhang, "The Entropy and PCA Based Anomaly Prediction in Data Streams," Procedia Computer Science, vol. 96, no. 1, pp. 139-146, 2016.
[41] "NUST," [Online]. Available: http://wisnet.seecs.nust.edu.pk/downloads.php. [Accessed 13 06 2013].
[42] "MAWI Working Group Traffic Archive," [Online]. Available: http://mawi.wide.ad.jp/mawi/. [Accessed 13 10 2016].
[43] A. E. Kossovsky, Benford's Law: Theory, the General Law of Relative Quantities, and Forensic Fraud Detection Applications, NewYork: WorldScientific, 2014.
[44] C. E. Shannon, "A mathematical theory of communication," Bell System Technical Journal, vol. 27, no. 4, pp. 623-656, 1948.

**Ali Naghash Asadi** is a Ph.D. candidate of computer engineering at Iran University of Science and Technology, Tehran, Iran. He obtained his B.Sc. degree in computer engineering from Guilan University in 2013, and his M.Sc. degree in computer engineering from Iran University of Science and Technology in 2015. His research include network traffic analysis, Petri nets, power and performance modeling, and stochastic and analytical modeling.

**Mohammad Abdollahi Azgomi** received the B.S., M.S. and Ph.D. degrees in computer engineering (software) (1991, 1996 and 2005, respectively) from Department of Computer Engineering, Sharif University of Technology, Tehran, Iran. His research interests include Petri nets, hybrid systems, stochastic modeling, quantitative evaluation, and trustworthy computing. Dr. Abdollahi Azgomi is currently an associate professor at School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.

# An Improved Sentiment Analysis Algorithm Based on Appraisal Theory and Fuzzy Logic

Azadeh Roustakiani
Department of Social Sciences and Economics, Alzahra University, Tehran, Iran
a.kianie@gmail.com
Neda Abdolvand*
Department of Social Science and Economics, Alzahra University, Tehran, Iran
n.abdolvand@alzahra.ac.ir
Saeedeh Rajaee Harandi
Department of Social Science and Economics, Alzahra University, Tehran, Iran
Rajaeeharandi.Saeedeh@gmail.com

Received: 24/02/2017 Revised: 05/05/2018 Accepted: 22/07/2018

**Abstract**

Millions of comments and opinions are posted daily on websites such as Twitter or Facebook. Users share their opinions on various topics. People need to know the opinions of other people in order to purchase consciously. Businesses also need customers' opinions and big data analysis to continue serving customer-friendly services, manage customer complaints and suggestions, increase financial benefits, evaluate products, as well as for marketing and business development. With the development of social media, the importance of sentiment analysis has increased, and sentiment analysis has become a very popular topic among computer scientists and researchers, because it has many usages in market and customer feedback analysis. Most sentiment analysis methods suffice to split comments into three negative, positive and neutral categories. But Appraisal Theory considers other characteristics of opinion such as attitude, graduation and orientation which results in more precise analysis. Therefore, this research has proposed an algorithm that increases the accuracy of the sentiment analysis algorithms by combining appraisal theory and fuzzy logic. This algorithm was tested on Stanford data (25,000 comments on the film) and compared with a reliable dictionary. Finally, the algorithm reached the accuracy of 95%. The results of this research can help to manage customer complaints and suggestions, marketing and business development, and product testing.

**Keywords:** Appraisal Theory; Fuzzy Logic; Sentiment Analysis; Opinion Mining.

## 1. Introduction

Recently, the sentiment analysis has become a very popular topic among computer scientists and researchers, because it has many usages in market and customer feedback analysis [1]. About 81% of Internet users search online at least once before purchasing a product, and 20% of them repeat it every day. Published reviews about products and services affect roughly 73% to 87% of buyers; hence, 20% to 99% of buyers prefer to choose the best services and products based on the existing opinions. These statistics indicate that customers now pay more attention to the views of other customers to receive more services and products. Customer opinions have always formed an important part of the information necessary for the decision-making. Before the advent of the World Wide Web (WWW), individuals asked the opinion of friends and experts to decide on purchasing products or services. However, the Internet and the Web have made it possible to understand the opinions and experiences of other people (regardless of the level of familiarity and expertise) [2]. With the advent of social networks, the possibility of interaction and communication between individuals has increased. This is because social networks have significantly tightened communication on the web and are being used by a wide range of people of different ages due to its' cheap, fast, and affordable access. The amount of data generated by web users during the exchange of information is also increasing. Individuals and companies that offer services or sell products have always been keen to see community feedback about their products and services [3].

Businesses need customer feedback to provide after-sales services, such as managing customer complaints, supporting and managing customer relationships, and predicting future sales. Therefore, sentiment analysis helps companies to know what customers think about their products so they can modify their products' features and introduce new products according to their customers' opinions [4]. People also need to know the opinions of other people in order to purchase consciously. Because, before buying, they will be aware of the experiences of other people and can decide which product is best. This requires the development of computational resources for the unlimited expansion of the expressing sentiment, as well as the increasing computing power to facilitate the processing of large amounts of data [5]. Extracting useful knowledge from this amount data is called sentiment

* Corresponding Author

analysis, which is widely used from business services to political campaigns [6]. With the development of social media, such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks, the importance of sentiment analysis has increased [4].

Studies that examine the feasibility of doing this with greater precision and lower costs have largely relied on two main methods. The first method is the "Bag of Words," which recognizes negative and positive documents based on the frequency of the occurrence of different words in the document; in this way, different learning methods can be used to select or weigh different parts of the text [4][7-8]. However, the performance of Bag of Words is somewhat limited due to a few fundamental deficiencies in handling the polarity shift problem [4]. Another method is called "contextual polarity," which usually divides words into good and bad and then calculates the good and bad points for the entire document [9-10]. However, these methods have ignored the important aspects of sentiment analysis. Therefore, a more precise contextual analysis of attitudinal expressions is required. Moreover, "separate units" of such expressions are not words but rather evaluation groups, such as "very good" and "not so funny", which express a particular tendency [10]. To date, few studies have incorporated data mining components into the fuzzy logic. Moreover, few researchers have focused on the semantic differences of the sentiment classification characteristics, with most tending to categorize texts with positive and negative sentiments. The positive or negative sentiments of the comments reduce the accuracy of the sentiment analysis algorithm. In fact, sentiments have several aspects that are fully addressed in the appraisal theory [11]. This theory is based on terms such as "very good" and "not so funny," in which an appraisal group is a set of specified values that are placed in several independent contextual classes and read the speaker or author' opinion. In fact, sentiment analysis provides a grammatical system for evaluating the writer or speaker's opinions [12].

Because of these problems and defects, this study, by combining the appraisal theory and fuzzy logic provides the third group of sentiment analysis methods for sentiment analysis and classification in the attempt to cover all the features of commented words and opinions.

This study begins with a review of the literature, followed by a description of the research method used and proposed algorithm. Finally, the implications and conclusion of the study will be explained.

## 2. Literature Review

Sentiment analysis is a text classification task that seeks to classify documents in accordance with the view of individuals (polarity) on a particular subject [13]. Sentiment analysis uses automated tools to detect subjective information, such as opinions, attitudes, and feelings expressed in text [4]. It includes numerous tasks, such as sentiment extraction and classification, mentality detection, and opinions summary [14]. The most active research area of sentiment analysis is natural language processing, which is widely studied in data mining, web mining, and text mining [4]. Pang and Lee [2] studied a variety of techniques and approaches that are used directly in sentiment-based information search systems and are used to convey a sense of excitement to the reader about the intellectual richness and extent of the area. Different studies use different sentiment analysis tools and techniques to improve the accuracy and quality of meta-analysis algorithms, and to achieve the three main objectives of managing customers' complaints and suggestions, marketing and business development, and product evaluation.

- **Manage Customers' Complaints and Suggestions**

All businesses realize that products and services must be endorsed by customers in order to ensure their business continuity. Before the advent of the WWW, the reactions of people to products depended on the people around them. But today, with the advancement of technology, people's perceptions of products or services are very fast, which leads to the failure or success of those products or services. Businesses can follow these published comments to easily identify their weaknesses and strengths, so, they can modify their products features [4]. Therefore, different methods are used in research of sentiment analysis. For example, Whitelaw et al. [10] provided a method for classifying sentiments using appraisal groups, in which all the limited features of reviews combined together by using the Bag of Words model and trying to manage customer feedback. Kanade et al. [4] have also used the Bag of Words model along with dual sentiment analysis to classify the reviews. The proposed system uses a dictionary-based classification for accurately classifying the reviews as positive, negative and neutral. Both the product owner and the user can identify the quality of the product based on the sentiment graph that is generated based on the reviews for each product. Pak and Paroubek [7] used linguistic analysis for sentiment analysis. They indicated how to automatically use a collection of writings for sentiment analysis and opinion mining objectives. Their sentiment classifier can determine positive, negative and neutral sentiments for a document. Moghaddam [15] have also proposed a technique to automatically extract defects and improvements from customer feedback for summarizing them. The results of this study indicated that without any manual annotation cost, the proposed semi-supervised technique can achieve comparable accuracy to a fully supervised model in identifying defects and improvements.

- **Marketing and Business Development**

In the past, word-of-mouth advertising was used for marketing and business development. But today, with the development of social networks, it is done at a faster pace and with more quality and trust. Businesses use big data and sentiment analysis techniques to advertise and expand their markets. This has led to the emergence of products and services' platforms for customers. In this regard, Li et al. [16] have proposed a recommender system based on

opinion mining to extract opinion related information from the massive reviews. They analyzed the linguistic information and designed a two-layer selection algorithm to find the most suitable products for customers. Their method had great accuracy, feasibility, and reliability.

- **Product Evaluation**

Evaluating the reaction of customers to the beta version of some products or services can be costly or, in some cases, impossible for reasons such as speed of product delivery to the market or royalty, which would cost business owners. With the introduction of analytical techniques, it is possible to assure customer's feedback to the product before it is delivered to the market. For example, Denecke and Deng [17] used sentiment analysis techniques to test their medical products and patient feedback. They performed a quantitative assessment with respect to word usage and sentiment distribution of a dataset of clinical narratives and medical social media, characterized the facets of sentiment in the medical sphere and identified potential use cases. Andreevskaia and Bergler [18] also presented a method for extracting fuzzy sentiments from WordNet using the Sentiment Tag Extraction Program (STEP). They indicated that the Net Overlap Score can be used as a measure of the words' degree of membership in the fuzzy category of sentiment.

There are several methods of sentiment analysis, one of which is appraisal theory. This theory is based on terms such as "very good" and "not so funny". Khoo et al. [11] used appraisal theory to determine the positive, negative and neutral sentiments of manual and automatic text news. Korenek and Šimko [3] also improved the speed and accuracy of the sentiment analysis algorithm using appraisal theory.

The presentation of a new method for classifying sentiments using appraisal groups goes beyond the categories of "positive" and "negative" [10]. Sentiment analysis with fuzzy logic due to its reasoning (and a closer look at exact sentiments) helps producers or consumers, or any other interested person, to make an effective decision regarding their favorite product or service [19]. Yadav et al. [20] presented a refined method for classifying keywords based on their sense relative to other keywords in the text. Fuzzy logic is used to classify these words expressing sentiments according to their application in the sentence. Dragoni et al. [13] has also used fuzzy logic to create the relationships graph in order to represent the appropriateness between sentiment concepts and different domains. They developed a semantic resource based on the connection between an extended version of WordNet, SenticNet, and ConceptNet, that has been used both for extracting concepts and for classifying sentences within specific domains. Krishna, Pandty, and Kumar [21] developed a new model for opinion mining and sentiment analysis, which uses the machine learning and fuzzy approach to classify the sentiment on textual reviews, and to automate the process of mining attitudes, opinions and hidden emotions from text. Apple et al. [22] proposed a hybrid classification model based on fuzzy sets, a solid sentiment lexicon, traditional NLP techniques and aggregation methods to investigate and devise solutions to the Sentiment Analysis Problems. They indicated that their hybrid method is much better at the sentence level. Keith et al. [14] used a hybrid approach that combines an unsupervised machine learning algorithm along with a natural language processing technique to analyze the reviews and to tag part of a speech (POS) to obtain the syntactic structure of a sentence. The syntactic structure, along with the use of dictionaries, can determine the semantic orientation of the reviews through an algorithm. Haseena Rahmath [23] also proposed a multi-step opinion mining system that involves pre-processing to clean the document, a rule-based system to extract features and a scoring mechanism to tag their polarity. The proposed technique utilizes fuzzy functions to emulate the effect of various linguistic hedges such as dilators, concentrator and negation on opinionated phrases that make the system more accurate in sentiment classification and summarization of users' reviews.

The prior studies are categorized into the three categories to identify the various tools for sentiments analysis and their details. Figure 1 demonstrated how these categories are related to the main output of this research:

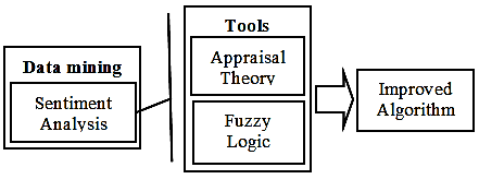


Fig. 1. Relationship between previous studies and present research

As can be seen from previous research, the studies have used two theories of appraisal theory and fuzzy logic separately to analyze the sentiments, but they have not studied the hybrid performance of these methods. Hence this study used a hybrid method of appraisal theory and fuzzy logic to improve the sentiment analysis algorithm. Many studies have been done in the field of sentiment analysis, therefore, a large number of data sets are the same data set that the proposed algorithm of this study is tested on. The importance of this is that we can compare the results of this research with the results of previous studies.

## 3. Conceptual Framework

This study is constructive and the CRISP-DM methodology, which is one of the greatest analytical methods for data mining projects, is used in this study. The general framework of the study is based on the research by Alamsyah et al. [24]. This framework analyses the sentiments using the appraisal theory. However, changes to the algorithm result in an unclear appraisal theory. In this framework, by applying the appraisal theory, the characteristics of comments are categorized and by using the techniques of text mining

and fuzzy logic, the orientation of the opinions from the positive and the negative is changed to the fuzzy range.

The algorithm has been implemented using the Python programming language. According to the CRISP-DM model, it is imperative to identify the business and data. All the opinions published on social networks and online shopping bases platforms pertain to the statistical population of this study. The proposed algorithm has been tested on Stanford University data. The data is divided into two groups: (i) positive and (ii) negative and needs no further preparation because the data type is textual and is ready to be processed in the programming environment.

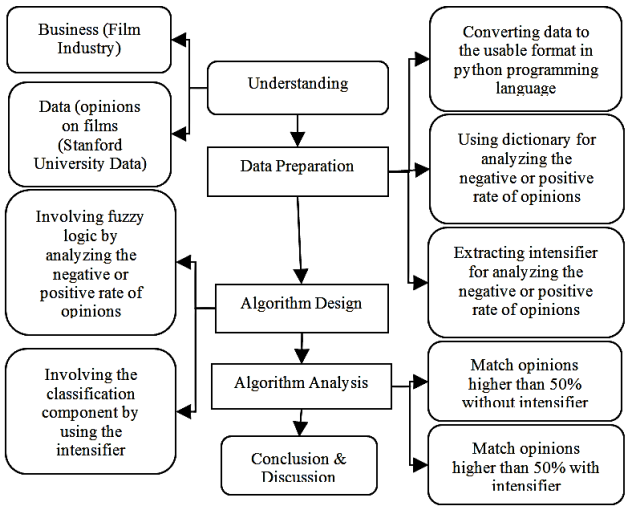


Fig. 2. Conceptual Framework

## 4. Algorithm Implementation

The Stanford University data used in this study comprise two sets of 12,500 negative and positive comments published about a movie [25]. These data are textual (txt) and their authors are anonymous. All the data are expressed in the author's own language and do not refer to anyone else. The importance of these comments can be examined in the following categories:

- Understanding customer's taste in the film
- Understanding the components of filmmaking for future releases

As all comments have been published in the author's own language, therefore, the discursive source constitutes the author's opinion. This dataset is divided into two groups: (i) Positive and (ii) negative. However, the degree of positive and negative is not known and thus would be determined through the proposed algorithm. Therefore, the fuzzy logic is used to determine the severity of the positive and negative aspects of the comments. To use fuzzy logic on data, it is necessary to have a dictionary that has positive and negative words, so that by comparing the dictionary and comments contained in the review of the movie, the degree of positive or negative comments will be determine. The majority of words in this dictionary have typos. In fact, these mistakes have been deliberately incorporated to include misleading words that are used extensively in social networks. In addition, a set of English adverbs has also been used to determine the classification and severity of the comment. For example, consider the following two sentences: the first sentence is positive at 87.5%. This figure is obtained comparing the positive sentence to the positive dictionary. As can be seen, there are seven positive words and a negative word, which results in a negative percentage of 12.5%. Each word that indicates the sentiment has a two-point score and each adverb has one point score thatits pseudocode is as below:

-------- file contents:

This show is awesome! I **love** all the actors! It has **great** story lines and characters. It is the **perfect** drama. James Caan and Josh Duhamel have **great** dialogue. They both can be really funny.I **miss** Vanessa Marcil on General Hospital, but she's **great** on here. James Lesure is **great**! He can be hilarious. Molly Sims plays a dimwit very well. The writing is awesome! They keep up an **excellent** pace. The show can really leave you hanging, which is one of my **favorite** elements of a show. I cannot wait until the new season starts. This show makes it to the **top** ten of all my shows. I hope this show stays on for a really long time. If people know what **good** is, it will. I never want the show to end. Ever.

-------- positive matches:

set(['perfect', 'great', 'good', 'love', 'top', 'favorite', 'excellent'])

-------- negative matches:

set(['miss'])

-------- results before matching the adverbs :

./data/positive/9201_10.txt - pos:14 - neg:2

./data/positive/9201_10.txt - pos:87.5 - neg:12.5

Subsequently, the classification and severity of the sentiments in the sentence is determined. This is done by assigning the coefficient to the sentence for each adverb. Because the adverbs represent the emphasis and strength of sentiments. Therefore, if one score is added to the total score of each positive clause, the effect of the adverb on the sentence will increase the positive percentage to 90%. Its peso code is as below:

-------- adverb matches :

set(['very', 'never', 'up', 'really'])

-------- results after matching the adverbs :

./data/positive/9201_10.txt - pos:18 - neg:2

./data/positive/9201_10.txt - pos:90.0 - neg:10.0

Table 1. Comparison of the Adverbs Effect in Sentiment Scoring

| Before the use of adverb | | After the use of adverb | |
|---|---|---|---|
| Positive | Negative | Positive | Negative |
| 14<br>87.5% | 2<br>12.5% | 18<br>90% | 2<br>10% |

The same trend is considered in negative statements that Its pseudocode is as below:

-------- file contents:

This movie is ridiculous. It's attempting to be a comedy but the screenplay is horrible. The whole movie is done in low light and you can't grasp the fact that it's a comedy. Truly is **bad** cinematography. You really have to sit there and watch it to realize there's a few jokes here and there going on but either way they're all inside jokes amongst themselves. This is more **like** a wannabe drama flick that went bad. It really is a very **pointless** movie. Their expressions reveal nothing but **dismay** and **disaster** which turns out that way anyway. Unless you want to be **bored** out of your ass, I suggest you stay away from this gag of a movie.

-------- positive matches:

set(['like'])
-------- negative matches:
set(['dismay', 'bad', 'bored', 'pointless', 'disaster'])
-------- results before matching the adverbs:
./data/negative/6533_1.txt - pos:2 - neg:10
./data/negative/6533_1.txt - pos:16.6666666667 - neg:83.3333333333

As it is clear, there are five negative words and a positive word that makes the sentence positive by 16.6%. Therefore, if one score added to the total score of each negative clause, the effect of the adverb on the sentence, will improve the negative percentage by about 5%. Its peso code is as below:

-------- adverb matches :
set(['very', 'away', 'there', 'really', 'more'])
-------- results after matching the adverbs :
./data/negative/6533_1.txt - pos:2 - neg:15
./data/negative/6533_1.txt - pos:11.7647058824 - neg:88.2352941176

Finally, the algorithm is run on all 25,000 comments on the movie browsing website. The results are shown in figure 3. As Figure 3 illustrates, the amount of negative words used in positive comments was greater than positive words, so that negative word strength overestimates the positivity of the opinion. By using the adverbs, the effect of the positive words in the positive comments increases and the accuracy of the algorithm in the calculation of the positive opinion increases. Figure 4 indicates the positive and negative outcomes of the comments after after including the adverbs in positive comments.

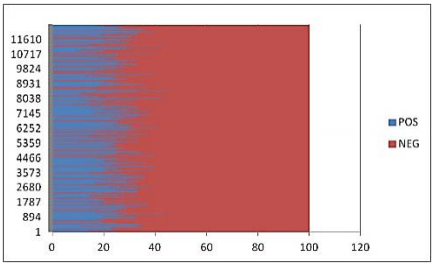


Fig. 3. The amount of positive and negative opinioon on positive comments

By using the adverbs, the effect of the positive words in the positive comments increases and the accuracy of the algorithm in the calculation of the positive opinion increases. Figure 4 indicates the positive and negative outcomes of the comments after after including the adverbs in positive comments.

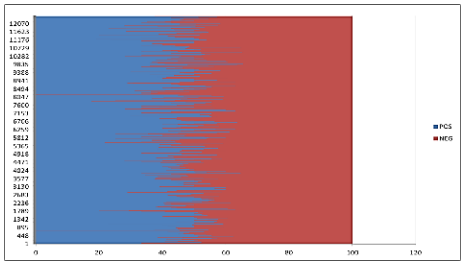

Fig. 4. The positive and negative outcomes of the comments after including the adverbs in positiv comments

The same is done for negative comments (Figure 5, 6). As shown in Figure 5, the negative words used in the negative comments were much higher than the positive ones.

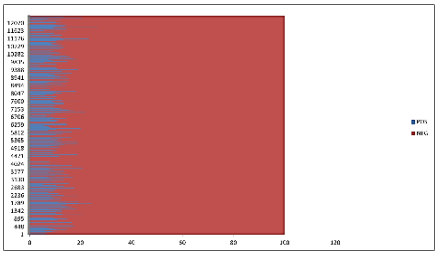

Fig. 5. The amount of positive and negative opinioon on negative comments

By using the adverbs, the effect of the negative words in the negative comments increases and the accuracy of the algorithm in the calculation of the positive opinion increases. Figure 6 indicates the positive and negative outcomes of the comments after including the adverbs in negative comments.

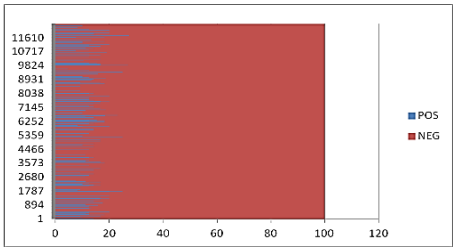


Fig. 6. The positive and negative outcomes of the comments after including the adverbs in negative comments

According to the results of the sentimen analysis, the movie, despite having positive comments, still has no general interest, and even positive comments contain a significant percentage of negative words. Moreover, the results indicate that before applying the appraisal theory, and only by using fuzzy logic, the positive feedback reached a precision of 79.632%, which improved after including the adverbs (the classification component of the appraisal theory) to 95.952%. Moreover, before applying the appraisal theory, and only by using fuzzy logic, the negative feedback reached a precision of 69.664%, which improved to 94.28% after including the adverbs (the classification component of the appraisal theory).

## 5. Comparison of the Proposed Algorithm with the Basic Algorithms

The accuracy of the proposed algorithm compare with the basic algorithms is indicated in table 2. According to the table 2, the study of Pang and Lee [8] divided comments into two positive and negative categories and achieved an accuracy of 86.4% in sentiment analysis. In addition, among the researches that have been analyzed the sentiments using the appraisal theory Whitewall et al. [10] earned the best score of 90% in the accuracy of sentiment analysis. The combination of fuzzy logic with appraisal theory has increased the accuracy of the sentiment analysis algorithm to 95%. In addition,

consideration of all the components of the appraisal theory as well as the fuzzy analysis of the opinions, and sentiment has improved the analysis algorithm. The results indicate a greater accuracy with the proposed algorithm compared to the previous algorithms (Ttable 2).

Table 2. The accuracy of the proposed algorithm compared with basic algorithms

| algorithm | Proposed algorithm | The algorithm of [8] | The algorithm of [10] |
|---|---|---|---|
| Accuracy | 95% | 86.4% | 90% |

## 6. Discussion and Conclusion

Internet and social media platforms resulted in changes not only to consumers' attitudes, perceptions and behaviours but also to the decision-making process itself. With the development of social media like as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks, the importance of sentiment analysis increased. Sentiment analysis is most active research areas in natural language processing which is widely studied in data mining, Web mining, and text mining. However, few studies have incorporated data mining components into the fuzzy logic. Moreover, few research have focused on the semantic differences of the sentiment classification characteristics. Therefore, this study, used a hybrid approach of the appraisal theory and fuzzy logic that has not been used in prior studies to cover all the features of commented words and opinions. Usually, the opinions about the film are divided into two categories of positive and negative. Therefore, the comments sentiment is quite clear. But the rate of positive or negative sentiments is not determined, that our proposed algorithm calculated the positive or negative percentage of words contained in the comment.

The adverbs used in the comment also determine the strength of the comment. In fact, the adverbs have been used to determine the classification of the appraisal theory. All the opinions expressed on the film are expressed in the author's own language. Therefore, the source of the commentary is the author. The study of Pang and Lee [8] has achieved an accuracy of 86.4% in sentiment analysis and divided comments into two positive and negative categories. In addition, among the researches that have been analyzed the sentiments using the appraisal theory Whitewall et al. [10] earned the best score of 90% in the accuracy of sentiment analysis. The combination of fuzzy logic with appraisal theory has increased the accuracy of the sentiment analysis algorithm. In addition, consideration of all the components of the appraisal theory as well as the fuzzy analysis of the opinions, and sentiment has improved the analysis algorithm. According to this study, before applying the appraisal theory, and only by using fuzzy logic, the positive feedback reached a precision of 79.632%, which improved after including the adverbs (the classification component of the appraisal theory) to 95.952%. Moreover, before applying the appraisal theory, and only by using fuzzy logic, the negative feedback reached a precision of 69.664%, which improved to 94.28% after including the adverbs (the classification component of the appraisal theory). The results indicate a greater accuracy with the proposed algorithm compared to the previous algorithms. The results of this study can be used for managing customer complaints and offers, sales forecasts, and the types of services and products available in the future, as well as for testing products and services in all customer-centric industries.

The use of the appraisal theory in sentiment analysis has increased in recent years. This method has three variables that should be identified and analyzed in the text. The study of these three variables is time-consuming and complex, requiring high computational power from the sentiment analysis machines. Therefore, the combination of this method with fuzzy logic was used only in the sentiment analysis, and the three other variables were not studied. Hence, it is suggested that future research develop an improved algorithm to analyze the users' sentiments in social networks based on the hybrid method of appraisal theory and fuzzy logic by considering all the variables of the appraisal theory.

In addition, the proposed algorithm involves fuzzy logic only in the trend orientation. However, fuzzy logic can be combined with all the appraisal theory variables to increase the accuracy of the model.

The algorithm used in this study can be used to measure the opinion and comments posted about various products. The output of this algorithm is usable for all products. The dictionary used in this research is a complete set of positive, negative and English constraints that cover a wide range of vocabulary. However, it is suggested to use this algorithm for comments posted in Persian by creating a Persian dictionary.

In addition, the proposed hybrid algorithm was validated on data from Stanford University, which was comments on a movie. These data were in English and had the language constraints. Therefore, it is suggested that future studies develop this algorithm in Persian-language social networks and evaluate its performance. Moreoer, In this research, we have been working with English, however, the proposed technique can be used with any other language.

The previous studies have used different data, which limits the possibility of comparing this method with previous methods. Morover, all the data of this research is quoted by the author him/herself, which practically influences the source of the discourse. Therefore, future research can, by using the classification variable in the appraisal theory examine the opinions expressed by others.

**Acknowledgement**

We would have pleasure to thank Mr. Mohammad Rafiee, for his technical expertise. He was so caring and ready to help regarding the manuscript. It's really hard to express our gratitude level to him for his immense contributions.

## References


[1] J. Fletcher and J. Patrick, "Evaluating the Utility of Appraisal Hierarchies as a Method for Sentiment Classification". Proceedings of the Australasian Language Technology Workshop 2005, (December 2005), pp. 134–142.

[2] B. Pang, and L. Lee, "Opinion Mining and Sentiment Analysis". Foundations and Trends in Information Retrieval, Vol. 1, No. 2, 2006, pp. 91–231.

[3] P. Korenek, and M. Šimko, "Sentiment analysis on microblog utilizing appraisal theory." World Wide Web, Vol. 17, No. 4, 2014, 847–867.

[4] M. A. Kanade, M. A. Deshmukh, M. S. Surwase, M.A. Kulkarni, and A. Zore, "Implementation on Intelligent Sentiment Review Analysis with Short Form Words and False Negative Comment Consideration". International Journal of Research in Engineering, Technology and Science, Vol. 3, No. 4, April 2018.

[5] J. Brooke, "A Semantic Approach to Automated Text Sentiment Analysis, 118." Retrieved from http://www.sfu.ca/~mtaboada/docs/Julian_Brooke_MA.2009.

[6] L. Zhang, and B. Liu, "Sentiment analysis and opinion mining. In Encyclopedia of Machine Learning and Data Mining" (pp. 1152-1161). Springer, Boston, MA. 2017.

[7] A. Pak, and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp. 1320–1326.

[8] B. Pang, and L. Lee, "A Sentimental Education: Sentiment Analysis using Subjectivity Summation based on Minimum Cuts". ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004, p. 271.

[9] T. Wilson, J. Wiebe, and P. Hoffman, "Recognizing contextual polarity in phrase level sentiment analysis". Acl, Vol. 7, No. 5, 2005, pp. 12–21.

[10] C. Whitelaw, C. Whitelaw, N. Garg, N. Garg, S. Argamon and S. Argamon, S. "Using appraisal groups for sentiment analysis". Proceedings of the 14th ACM International Conference on Information and Knowledge Management - CIKM '05, 2005, p. 625.

[11] C. S.-G. Khoo, A. Nourbakhsh, A., and N. Jin-Cheon, N. "Sentiment analysis of online news text: a case study of appraisal theory". Online Information Review, Vol. 36, No. 6, 2012, pp. 858–878.

[12] K. Bloom, K. Sentiment analysis based on appraisal theory and functional local grammars. Illinois Institute of Technology. ProQuest Dissertations Publishing, 2011. 3504518.

[13] M. Dragoni, A. G. Tettamanzi, and C. da Costa Pereira, (2014, May). "A fuzzy system for concept-level sentiment analysis." In Semantic web evaluation challenge (pp. 21-27). Springer, Cham.

[14] B. Keith, E. Fuentes, and C. Meneses, "A Hybrid Approach for Sentiment Analysis Applied to Paper" .In Proceedings of ACM SIGKDD Conference, Halifax, Nova Scotia, Canada, August 2017 (KDD'17), 2017, P. 10

[15] S. Moghaddam, "Beyond sentiment analysis: mining defects and improvements from customer feedback." In European Conference on Information Retrieval (pp. 400-410). Springer, Cham. 2015, March.

[16] H. Li, J. Ding, D. Nie, and L. Tang, "Accurate Recommendation Based on Opinion Mining". Advances in Intelligent Systems and Computing, Vol. 329, 2015, pp. 325–333.

[17] K. Denecke and Y. Deng, "Sentiment analysis in medical settings: New opportunities and challenges". Artificial Intelligence in Medicine, Vol 64, No. 1, 2015, pp. 17–27.

[18] A. Andreevskaia, and S. Bergler, "Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses". Proceedings of EACL, 6, 2006, pp. 209–216.

[19] T. Rahman, "Sentiment Analysis by Using Fuzzy Logic", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 4, No. 1, 2014, pp. 33–48.

[20] D. K. Tayal, and S. K. Yadav, "Word level sentiment analysis using fuzzy sets". Analysis, Vol. 54, 2015, p. 59.

[21] B. V. Krishna, A. K. Pandey, and A. S. Kumar, "Feature Based Opinion Mining and Sentiment Analysis Using Fuzzy Logic". In Cognitive Science and Artificial Intelligence (pp. 79-89). Springer, Singapore. 2018.

[22] O. Appel, F. Chiclana, J. Carter, and H. Fujita, "A hybrid approach to the sentiment analysis problem at the sentence level". Knowledge-Based Systems, Vol. 108, 2016, pp. 110-124.

[23] P. Haseena Rahmath. "Fuzzy based Sentiment Analysis of Online Product Reviews using Machine Learning Techniques." International Journal of Computer Applications (0975 – 8887). Vol. 99, No. 17, 2014, pp. 9-16

[24] A. Alamsyah, W. Rahmah, and H. Irawan, "Sentiment analysis based on appraisal theory for marketing intelligence in Indonesia?" Mobile phone market. Journal of Theoretical and Applied Information Technology, Vol. 82, No. 2, 2015, pp. 335–340.

[25] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis". Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011. pp. 142–150.



**Azadeh Roustakiani** holds a Master's in Information Technology Management from the Alzahra University, Tehran, Iran. She is interested in research in the field of information systems and technology, including innovation in ICT, business intelligence, electronic commerce, and big data analytics.

**Neda Abdolvand** is an Assistant Professor in the Alzahra University and was a Postdoctoral Research Fellow at Tarbiat Modares University. She holds a PhD and MS in Information Technology from the Tarbiat Modares University and a Postgraduate Certificate in Information Systems from Melbourne University. She is interested in research in the field of information systems and technology, including innovation in ICT, business intelligence, electronic commerce, and big data analytics. She has been a member of the Association for Information Systems since 2009.

**Saeedeh Rajaee Harandi** holds a Master's in Information Technology Management from the Alzahra University, Tehran, Iran. She is interested in research in the field of information systems and technology, including innovation in ICT, business intelligence, electronic commerce, and big data analytics.

# Toward Energy-Aware Traffic Engineering in Intra-Domain IP Networks Using Heuristic and Meta-Heuristics Approaches

Mojtaba Sabahi Aziz
Computer Engineering Department, Bu-Ali Sina University, Hamedan, Iran
s3b3hi@gmail.com
Sepideh Zarei
Computer Engineering Department, Bu-Ali Sina University, Hamedan, Iran
sepideh.zarei08@gmail.com
Muharram Mansoorizadeh*
Computer Engineering Department, Bu-Ali Sina University, Hamedan, Iran
mansoorm@basu.ac.ir, muharram@gmail.com
Mohammad Nassiri
Computer Engineering Department, Bu-Ali Sina University, Hamedan, Iran
m.nassiri@basu.ac.ir

Received: 02/06/2018 Revised: 20/08/2018 Accepted: 21/08/2018

**Abstract**

Because of various ecological, environmental, and economic issues, energy efficient networking has been a subject of interest in recent years. In a typical backbone network, all the routers and their ports are always active and consume energy. Average link utilization in internet service providers is about 30-40%. Energy-aware traffic engineering aims to change routing algorithms so that low utilized links would be deactivated and their load would be distributed over other routes. As a consequence, by turning off these links and their respective devices and ports, network energy consumption is significantly decreased. In this paper, we propose four algorithms for energy-aware traffic engineering in intra-domain networks. Sequential Link Elimination (SLE) removes links based on their role in maximum network utilization. As a heuristic method, Extended Minimum Spanning Tree (EMST) uses minimum spanning trees to eliminate redundant links and nodes. Energy-aware DAMOTE (EAD) is another heuristic method that turns off links with low utilization. The fourth approach is based on genetic algorithms that randomly search for feasible network architectures in a potentially huge solution space. Evaluation results on Abilene network with real traffic matrix indicate that about 35% saving can be obtained by turning off underutilized links and routers on off-peak hours with respect to QoS. Furthermore, experiments with GA confirm that a subset of links and core nodes with respect to QoS can be switched off when traffic is in its off-peak periods, and hence energy can be saved up to 37%.

**Keywords:** Energy-aware Traffic Engineering; Green Networking; Greedy Algorithms; Genetic Algorithms.

## 1. Introduction

The Internet is expanding very fast. Reports generated in 2007 indicate that about 5.5% of whole world energy consumption is related to the Internet, and this number is annually increased by the rate of 20-25% [1]. In 2012 American's home equipment such as modems, routers, and gateways consumed about 803 TWH electricity and produced five million tons of $CO_2$.

Efforts toward power consumption management in networks cover a wide variety of methods ranging from energy-proportional routing to moving data centers to geographical locations that offer low-cost and/or nature-friendly electricity. Recently green networking solutions [2] were presented with the aim of reducing $CO_2$ and energy cost by designing energy-aware protocols and planning and manufacturing low power devices.

In traditional networks, routing takes place with static parameters in which packets of data select shortest paths to their destination. Traffic engineering is the task of routing network traffics in an efficient and reliable manner. It uses link load variations as routing parameters in building paths with the aim of optimizing an objective function. Basically, IP routing protocols try to route traffics over shortest paths, considering link weights as a metric. However, focusing on the optimal path may lead to congestion on links constituting the path. Traffic engineering methods try to re-shape and balance the load so that links and paths in other parts of the network are also involved. The main objective here is to keep link utilization in predefined bounds [3]. Energy-aware traffic engineering (EATE) extends the concerns to new dimensions related to power consumption, its costs, and side-effects [4].

As a subfield of green networking, EATE considers energy consumption in routers as the main parameter in routing decision. According to recent statistics, link utilization in networks providing Internet is about 30-40% [5]. Although green networking has attracted lots of attention in recent years, few works have been done on this subject [6]. Low utilized network links, lack of preferable energy management methods for network infrastructures,

* Corresponding Author

increasing cost of energy, increasing number of Internet subscribers, and increasing number of ISPs, are motivations for developing energy-aware traffic engineering approaches.

Turning off network elements to save energy is a key insight in developing energy-aware solutions; however, selection of the elements to turn off is not a trivial task. For a fixed topology and known traffic demand, several subsets of the network elements can be candidates for deactivation. Identifying the best candidate is computationally very expensive, and hence is not feasible for practical applications. Looking for feasible sub-optimal solutions by using approximation algorithms such as greedy approaches, search techniques enriched by heuristics, and random search techniques would be a natural decision for this problem.

Another consideration in the decision for turning off an element is its effect on network stability. For a given traffic demand, an approach can nominate a low utilized link for removal but the link may be required in a slightly modified traffic pattern. These subsequent changes to the network architecture make it unstable and produce management overhead. Hence, any modification to the network architecture must take future traffic variations into account. Estimating future traffic pattern can help in selecting suitable removal candidates. The effectiveness of the approach is directly related to the precision of the estimation process.

In the context of intra-domain traffic engineering, this paper proposes four algorithms for efficient tailoring of the network. The algorithms turn off a subset of network elements and produce a set of paths to transfer the demanded traffic. Selection procedure considers both energy consumption and network stability. Since there are multiple paths from a source to a destination and the paths are usually low-utilized, their traffics can be aggregated and routed over a single path. Elements of the other paths can be deactivated to save energy. It is important to notice that deactivation of links and nodes should not degrade network performance metrics such as maximum link utilization, packet delay, and reliability.

Our main contribution is twofold. First, our approach separates topology management from the routing decision. In this step, we do not alter settings of route setup procedures, link weights, and hyper-parameters. Secondly, we propose a set of new heuristics for sleeping nodes and links. The heuristics are based on the out of the box information from routing infrastructure.

## 1.1 Preliminaries

In this section, energy usage model of network routers is described. Furthermore, since our approach is based on DAMOTE (Decentralized Agent for MPLS Online Traffic Engineering) algorithm [7], [8] we introduce the algorithm and emphasize its important features.

### 1.1.1 Energy Usage in Network's Hardware Components

A network router is the main hardware device for traffic engineering. Network routers are composed of a chassis and a number of line cards which have deactivation capability. Line cards are the most important energy users in a router. For instance, 43% of energy usage in a Cisco 12000 router[1] is in its line cards.

We assume that a router has a single line card and each line card has 12 ports. A network link connects two distinct router ports. Line cards consume 40 watts and each port, when it is idle, consumes two watts. Additional 1.73 watts is consumed when a port transmits data with its whole potency. The maximum power consumption of each port will be 3.37 watts. Every port in this paper has 1Gbps bit rate. Comprehensive analysis of network power consumption is discussed in [9]. When all ports of a line card are inactive, the whole line card will be deactivated to save more power. Power usage of a router is calculated by Eq. 1 [9].

$$\textbf{Eq. 1} \quad P_r = P_{ch} + N_{ln} \times P_{ln} + \sum_{i=0}^{K} (UF \times P_i \times \mathrm{NP}_i) + N_p \times C$$

Where, $P_r$ *is* total power usage in the router and $P_{ch}$ is power consumption in its chassis. $N_{ln}$ and $N_p$ indicate the number of line cards and ports respectively. $K$ is the number of different port configurations. $UF$ is port utilization factor and $C$ is the constant power usage of each port regardless of traffic crossing over it.

### 1.1.2 Network Energy Proportionality Index

In Figure 1 power consumed by a device is plotted against the load on the device (in Gbps or active number of ports). Ideally, the power consumed should be proportional to the load, with the maximum power consumed, M, being as low as possible. The *ideal* curve represents this desired behavior. In practice, the behaviour of network devices follows the *measured* curve in which the device consumes at least $I$ watts. The difference between ideal and measured curves forms the basis of the following energy proportionality index (EPI) for networking components:

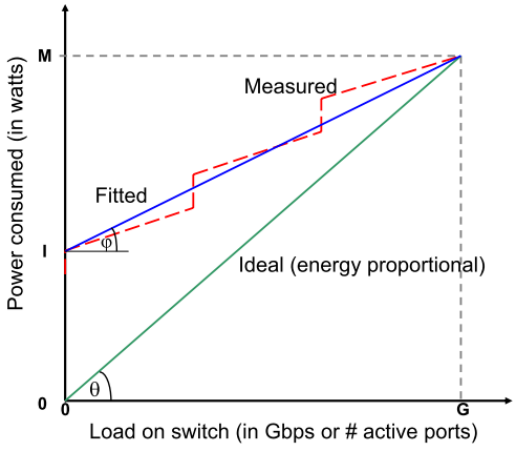


Fig. 1. Ideal and measured power consumption of network devices [9]

$$\textbf{Eq. 2} \quad \begin{aligned} EPI &= (M - I)/M \times 100 \\ EPI &= \tan\theta/\tan\varphi \times 100 \\ Normalized\ Power &= M/G \end{aligned}$$

1. http://www.cisco.com/c/en/us/products/routers/ 12000-series-routers/index.html

If θ and φ are the angles at the origin for the ideal and measured power, EPI is simply $(\tan(\varphi)/\tan(\theta)) * 100$. We express EPI in percentages, with 100 implying that the device has perfect energy proportionality and 0 implying that the energy consumed by the device is always a constant value. Note that EPI is independent of the maximum load that can be carried by the device. Thus, it is most useful in comparison of energy proportionality of devices in the same class. Furthermore, normalized power is the maximum power consumed by the device, *M*, divided by its aggregated bandwidth, *G*. Equation calculates the energy consumption of a router without taking into account the basic power usage, *I*. Actually, energy consumption of transferring one megabit is calculated by equation 3:

**Eq. 3** $$Energy_{1-mbps} = (M - I)/G$$

This value is used in estimating power consumption of a network topology serving a traffic demand.

### 1.1.3 The DAMOTE Algorithm

Basic routing function in this paper uses DAMOTE algorithm. DAMOTE is an MPLS routing function that sets up LSPs online and incrementally in order to optimize an objective function. It uses Bellman-Kalaba shortest path algorithm in building up routing paths.

Different objective functions have been used in DAMOTE of which we focus on the following one that combines load balancing and traffic minimization, as in Eq 4. Here, α allows tuning the trade-off between load balancing and traffic minimization. Lower $\alpha$ will favor longer paths and smooth the load throughout of the network, whereas a greater α will try to minimize the traffic over links. In the case of α = 0, it gives a low blocking probability by avoiding single point of failure. The ordered pair *(i,j)* indicates a link of the set *U*. $X^{aux}_{(i,j)}$ denotes the amount of reserved bandwidth on the link. $W^{cap}_{(i,j)}$ denotes capacity of the link and $\frac{\overline{X^{aux}}}{W^{cap}}$ is the average reserved capacity of all links.

**Eq. 4**
$$\sum_{(i,j)\in U}\left(\frac{X^{aux}_{(i,j)}}{W^{cap}_{(i,j)}} - \frac{\overline{X^{aux}}}{W^{cap}}\right)^2 + \alpha \sum_{(i,j)\in U}\left(\frac{X^{aux}_{(i,j)}}{W^{cap}_{(i,j)}}\right)^2$$
$$\text{with } \frac{\overline{X^{aux}}}{W^{cap}} = \frac{1}{|U|}\sum_{(i,j)\in U}\frac{X^{aux}_{(i,j)}}{W^{cap}_{(i,j)}}$$

The rest of this paper is organized as follows. Section two gives a brief overview of the related recent works. Section three presents motivation, and proposed algorithms for energy-aware traffic engineering. Analysis of the experimental results is presented in section four. Section five contains some concluding remarks and draws a few future directions.

## 2. Related Work

Energy-aware traffic engineering has been the subject of many works in recent years. For a thorough and extensive review of the literature see [2], [4], [10], and [11]. Existing methods can be grouped into three main categories, namely *rate adaptation*, *infrastructure sleeping*, and *green networking* [4]. Following the subject of this research, sleep based approaches are reviewed in more detail.

Earlier efforts tried to shed a light on theoretical aspects and proposed several performance measures feasibility analysis. Green-TE is the first energy-aware traffic engineering algorithm [12]. Its basic function is based on a centralized coordinator adopting traffic engineering decisions. In order to reach reliability, this coordinator can be replicated in different locations over the network. Coordinator's responsibility is to gather information from routers, solve the Green-TE problem, and obtain a configuration, and then distribute the decisions to the routers. Based on these decisions a router deactivates some or all of its ports. This method could reduce 27-42% of energy usage. In order to deal with failures, in [13] energy-aware mechanism uses two link-disjoint paths to route traffic demands.

Another approach is proposed in [14] for IP networks. This approach deactivates line cards and chassis of the router to save energy. It consists of three main phases. In the first phase traffic matrices are sorted in decreasing order of the whole traffic demand. The second phase tries to reduce energy consumption by solving a linear optimization problem. The last phase manages congestion while preserving energy consumption within the predetermined bounds.

The approach proposed in [15] is an energy-aware solution for backbone networks. They try to turn off an unused subset of line cards constituting a logical link. The approach could reduce energy consumption by 79% in some cases.

Amaldi et al. [16] modeled and discussed the energy-aware routing problem as a *Mixed Integer Linear Programming (MILP)* optimization. A deterministic solution for this problem can be found by search techniques such as backtracking and branch-and-bound. However, these techniques may require an exponential amount of time and memory in worst case scenarios. Since the problem is NP-hard, they proposed heuristic approaches based on *Interior Gateway Protocol Weight Optimization (IG-WO)* to find feasible sub-optimal solutions. Their approach is composed of greedy search steps that exploit properties of IG-WO.

Ruiz-Rivera et al. [17] Studied the problem of reducing energy consumption in MPLS networks. They compared online and offline heuristics for LSP setup; concluding that for well-known topologies such as Abilene and AT&T it is desirable to achieve LSP acceptance rates above 90% with up to 20% of links shut down.

Fortz reviews meta-heuristic approaches for traffic engineering in IP networks [18]. The studied works mostly try to set link weights in OSPF based routing. The goals of these works are controlling congestion and finding low-cost routes. Few of the studied works also consider node failures and link breakdowns. Compared to the context of energy-aware routing, these goals are short-term and highly dynamic. While the mentioned works proved to be efficient in terms of finding proper routes, their real-time application remains a challenge.

An approach based on genetic algorithms (GAs) is proposed in [19] to reduce network energy consumption besides reducing network reconfiguration rate. GA is a meta-heuristic algorithm in which an initial generation evolves by passing through some generations. Every member in each generation keeps topology information. A member is feasible when it can transfer all the traffic demands and at the same time satisfy maximum power consumption constraint. A member's fitness value is the weighted sum of normalized power consumption and reconfiguration costs. Samadi et al. [20] adopted *Non-dominated Sorting Genetic Algorithm II (NSGA-II)* for load balancing and energy consumption management. Their multi-objective approach induces sub-optimal Pareto-fronts as solutions.

In summary, existing works model the energy-aware traffic engineering as NP-hard mixed integer programming problem, and hence try to induce near-optimal solutions by means of various heuristics and metaheuristics. The approaches vary from self-sleeping routers to inter-layer energy-aware protocols. However, there is a strong agreement that to be applicable in practical scenarios, a solution must be used on top of existing OSPF and MPLS based networks. As an important challenge, solutions that exploit protocol-specific features (e.g. link weights in route setup procedure) are tightly coupled to the special protocol that limits their application and flexibility. For example, a solution that is developed on IGP-WO link weights needs to be redesigned for MPLS networks. Furthermore, its modifications to link weights may interfere with other network tasks such as QoS enforcement operations.

## 3. Proposed Approach

Theoretically, a network with $M$ nodes and $N$ links may have $2^{M+N}$ different topologies. As discussed in related works, (e.g. [2] and [16]), Finding the best topology in this huge space is an NP-hard problem. Hence, instead of trying to find the optimal solution, one can try to find feasible sub-optimal solutions in a reasonable amount of time and computation resources. Various techniques have been developed for this sub-optimal searching problem. These techniques are broadly classified into three main categories:

- **Graph Traversal:** graph traversal algorithms model the solution space as a graph that its nodes are candidate solutions. There is a link between nodes A and B if solution A can be transformed to B by a set of pre-defined operations (e.g. by flipping order of nodes). Backtracking and Branch-and-Bound are well-known search algorithms in this category. The former traverses the graph by depth-first order and the later follows breadth-first order. In a worst-case, backtracking may fail after an exponential amount of time without reporting a feasible solution. Similarly, branch-and-bound search may require an exponential amount of memory for queuing intermediate solutions. Due to these limitations, graph traversal algorithms are not suitable for practical scenarios.
- **Greedy Search:** greedy algorithms build solutions gradually by taking locally best steps. Naturally, these algorithms are deterministic and require a polynomial amount of time and memory. The drawback is that greedy algorithms usually converge to local minima that can be far away from optimal solutions.
- **Random search:** Random search techniques try to inspect diverse parts of the solution space. A random search procedure starts with an initial point in the solution space and then iteratively generates new solutions in the neighborhood of the current solution. Genetic algorithms (GA's) are a type of random search technique that combines random selection with biological evolution concepts. For discrete selection problems (as in our case of topology selection) it is known that GA's are better suited than their counterparts such as Particle Swarm Optimization (PSO) based techniques [21].

We assume that traffic demands are stable and smooth at least for a few hours. That is, we do not consider large traffic fluctuations in short time intervals. Because frequent modifications of network structure by turning elements on and off will make it very unstable that is an important challenge by itself and must be addressed separately [22], [23]. However, we try to compensate for small variations by limiting maximum link utilization.

Our main approach is to generate candidate topologies out of available network and then select the best one for traffic delivery. We treat the underlying routing procedure as a black-box element that accepts a network structure and a traffic demand matrix, then tries to route the demand over the network.

As depicted in Figure 2, our approach has two distinct modules. The first module generates topologies by means of heuristic and meta-heuristic algorithms. The evaluation module feeds the generated topology along with the demanded traffic to the routing procedure and grabs its output. The routing procedure outputs routing information such as built paths, employed links, and their occupied capacities. The module then analysis these information and estimates feasibility and usefulness of the topology.

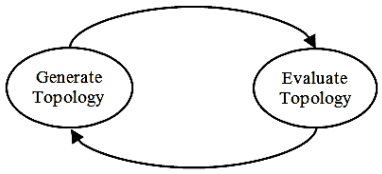


Fig. 2. Main modules of the proposed approach

Based on greedy and random search techniques, we propose four algorithms to employ energy-awareness in routing. The rationale behind these approaches is that simpler networks, e.g. networks with small a number of links and nodes, will consume less energy. Hence, we try to turn off ports, links, and nodes as much as possible. It is quite obvious that the simplified network must serve the demanded traffic as the original network.

### 3.1 Sequential Link Elimination

Sequential Link Elimination (SLE) is a greedy backward selection procedure (Algorithm 1) that switches off links one by one and evaluates the resulting network. If the network can transfer the demanded traffic and overall utilization is lowered, the link will be marked as a candidate for inactivation. After processing all of the links, the candidate with minimum utilization will be switched off. This process is repeated until no more links can be deactivated. Finally, nodes with no active links will be turned off to save more energy.

Algorithm 1. Sequential Link Elimination(SLE)

```
1.  Input: Original Network(V,E),
    Traffic Demand (D)
2.  Output: Prunned network
3.  finished = false
4.  repeat
5.    for e ∈ E
6.      util[e] = max_util(<V,E- e>)
7.    e = argmin(util)
8.    if util[e] < threshold
9.      E = E - e
10.   else
11.     finished = true
12. until finished = true
13. for v ∈ V
14.   mark[v] = false
15. for <u, v> ∈ E
16.   mark[u] = true
17.   mark[v] = true
18. for v ∈ V
19.   V = V-v iff mark[v] = false
20. return (<V, E>) // Reduced network
```

#### 3.1.1 Time Complexity

The main operation of this algorithm is calling DAMOTE. Suppose that the network under consideration has *N* active links and *M* nodes. In each iteration, the algorithm will deactivate a link or terminate. In worst case, DAMOTE is called $N + (N-1) + \ldots + 1$ times. which belongs to $O(N^2)$. Combined with DAMOTE's running time, O(NM), total complexity of the algorithm is $O(N^3M)$.

A simple improvement to the SLE would be applying *maximum utilization constraint* that does not allow links to use more than 80 percent of their capacity. This constraint prevents turning off some links and leads to less energy saving but more robust topology against demand variations.

### 3.2 Extended Minimum Spanning Tree

For an undirected graph, a spanning tree is a tree composed of the graph's nodes and a subset of its links that preserves connectivity of the original graph. Assuming that each link of the graph has a non-negative real-valued weight, cost of a tree is the sum of weights of its links. A minimum spanning tree (MST) is a spanning tree with the minimum cost. Algorithms such as Prim and Kruskal efficiently build MSTs [24].

An immediate idea for pruning network would be replacing the original network with its MST. The weight of a link connecting two nodes is the minimum number of active ports of these nodes.

A problem of this algorithm is its inability to turn off nodes. Extended minimum spanning tree (EMST) (Algorithm 2) resolves this issue by detecting removable nodes. Suppose that a path is established for a given traffic demand. Source and destination of the path are called *edge* nodes. Other nodes in the path are *core* nodes. In this algorithm, shortest paths between all pairs of edge nodes are calculated. Nodes which are not in any shortest path will be switched off. Core nodes are then sorted according to their frequency of presence in the shortest paths. The node with the lowest presence will be turned off and the shortest paths between every pair of edge nodes will be re-established. If there is a path for each pair of the edge nodes, the algorithm goes on with the next core node in the list. But, if there is a pair with no path, the core node is turned back on and the algorithm terminates. The final topology excluding inactive nodes and links is fed to the minimum spanning tree algorithm.

The topology must be able to transfer traffic demand. If it fails, some nodes and links should be added to it as the post-processing step. This step sequentially adds links to the topology and terminates when the topology can transfer the demand.

Algorithm 2. Extended Minimum Spanning Tree (EMST)

```
1.  Input: Original Network(V,E),
    Traffic Demand (D)
2.  Output: Pruned network
3.  P = build_path(<V , E>, D )
4.  for v ∈ V
5.    f[v] = #{p|p∈P ∧ v∈p}
6.  for v ∈ V
7.    V = V – v if f[v]= 0
8.  descend_sort(f)
9.  for each v in f
10.   disable (v)
11.   if isvalid (<V-v , E>, D)
12.     T = MST( <V-v , E> )
13.     if isvalid(T, D)
14.       V = V - v
15.     else
16.       exit for
17. return <V,E> //pruned network
```

#### 3.2.1 Time Complexity

Suppose that M and N are the number of edge and core nodes, respectively. Using Floyd's all-pairs-shortest-path algorithm, the time complexity of finding shortest paths is $O(M+N)^3$. $M^2$ is the number of all possible paths. Since each core node is searched in all of the $M^2$ paths and maximum path length is N, searching is of order $O(M^2N)$. Sorting the list of core nodes takes O(Nlog(N)) time and another $O(M^2)$ time is required for minimum spanning tree construction. Summing up the terms, the algorithm requires $O(M+N)^3$ running time.

### 3.3 Energy-aware DAMOTE

Energy-aware DAMOTE (EAD) switches off elements of network infrastructure to decrease topology size and save energy and turns them back on when they are required.

The objective of EAD is minimizing maximum link utilization and minimizing energy consumption of network.

At first, DAMOTE algorithm is run on the topology and link utilizations are obtained. Links with zero utilization are then turned off. These links are redundant and can be used in the case of failures. Other links are sorted according to their utilization and energy priority and then sequentially processed for removal (Algorithm 3).

Links are first sorted according to their utilization in a list, $L_1$, and also according to their energy priority in another list, $L_2$, in ascending order. The relative order of a link in the combined list, $L$, is determined by sum of its positions in $L_1$ and $L_2$. For instance suppose that for a network with four links, namely {e1, e2, e3, e4}, relative order of utilization is {e1, e2, e4, e3} and relative order of energy is {e4, e3, e2, e1}. Taking into account both factors, the combined order would be {e3, e1, e4, e2}.

Algorithm 3 Energy-aware DAMOTE (EAD)

```
1.  Input: Original Network(V,E),
    Traffic Demand (D),
    MLU threshold (a)
2.  Output: Pruned network
3.  A = evaluate (<V,E> , D) ;
4.  L1 = sort_by_energy(A, E)
5.  L2 = sort_by_util (A, E)
6.  L = combine(L1,L2)
7.  for e ∈ L
8.    A = evaluate (<V,E –e> , D) ;
9.    if isvalid(A)
10.     if max_util(A) < a
11.       E = E –e //permanently
12.     else
13.       enable(e)
14.   else
15.     exit for
```

Each link is marked for removal and the topology excluding it is tested against the traffic matrix. If it can handle the traffic, the link is turned off. Furthermore, nodes with no active links are turned off to save more energy.

A similar argument as EMST applies for EAD, concluding that time complexity of this algorithm is also at most $O(|V|+|E|)^3$

## 3.4 Genetic Algorithms

Genetic algorithms (GAs) are tools for random search in potentially diverse and huge solution spaces [25]. In GA terminology, a solution is called *an individual*. Basically, a GA procedure starts with an initial set of candidate solutions that is called the *first generation*. Each new generation is induced from earlier generations with the aid of genetic operations. The operations combine multiple individuals or modify a single individual to get better individuals. While GA does not guarantee to find an optimal solution, it usually finds a feasible and good sub-optimal solution.

Formulation of an optimization problem to be solved by GA is composed of several steps. At first, genes and chromosome structures must be built up. Then, procedures for initiating the first generation, evaluation of solutions and induction of subsequent generations must be defined. In the literature, usually, classic GA procedures are adopted with a few extensions for GA operations. In the following subsections, details of using GA for topology pruning is presented.

### 3.4.1 Genes and Chromosomes

As depicted in Figure 3, a network with M nodes and N links is represented as a chromosome with M+N genes. In this formulation, a gene is a three-state variable denoting presence or absence of the respective node or link. A zero-valued gene denotes that the respective node or link is inactive (disabled or turned off) so that it does not consume energy. A gene with a value of one denotes that the respective node or links is active and consumes energy. A value of two for a gene denotes that corresponding node or link must be permanently active. Permanent nodes and links are identified automatically based on the demanded traffic. That is, source and destination nodes of all the flows are marked as permanent nodes. If a permanent node is connected to the network with just a single link, the link is also marked as permanent to enforce network connectivity.

### 3.4.2 Initial Population

After marking permanent nodes and links, the initial population is generated by random setting of non-permanent genes. This stage requires population size which is usually set by the user.

### 3.4.3 Evaluation

Usefulness or *fitness* of a solution is proportional to its energy saving. Eq. 5 calculates fitness value; *f;* for a given solution; x. $E_{old}$ is the energy that is consumed by the basic topology. $E_{new}$ is the energy consumption of the modified topology. Solutions that fail to service all the demanded traffic or violate maximum utilization constraint are ignored.

**Eq 1.**
$$f(x) = \begin{cases} 1 - \left(\frac{E_{new}}{E_{old}}\right), & if \max_{util} < 0.8 \\ 0, & otherwise \end{cases}$$

### 3.4.4 Selection and Reproduction

The concept of evolution in GA starts with the *selection* procedure in which, two members of the current generation are selected for reproduction. These members are called parents and the newly generated individuals are called *off-springs* or *children*; analogous to reproduction procedure in most biological systems. We use well-known roulette-wheel selection method that assigns selection probabilities proportional to its fitness values.

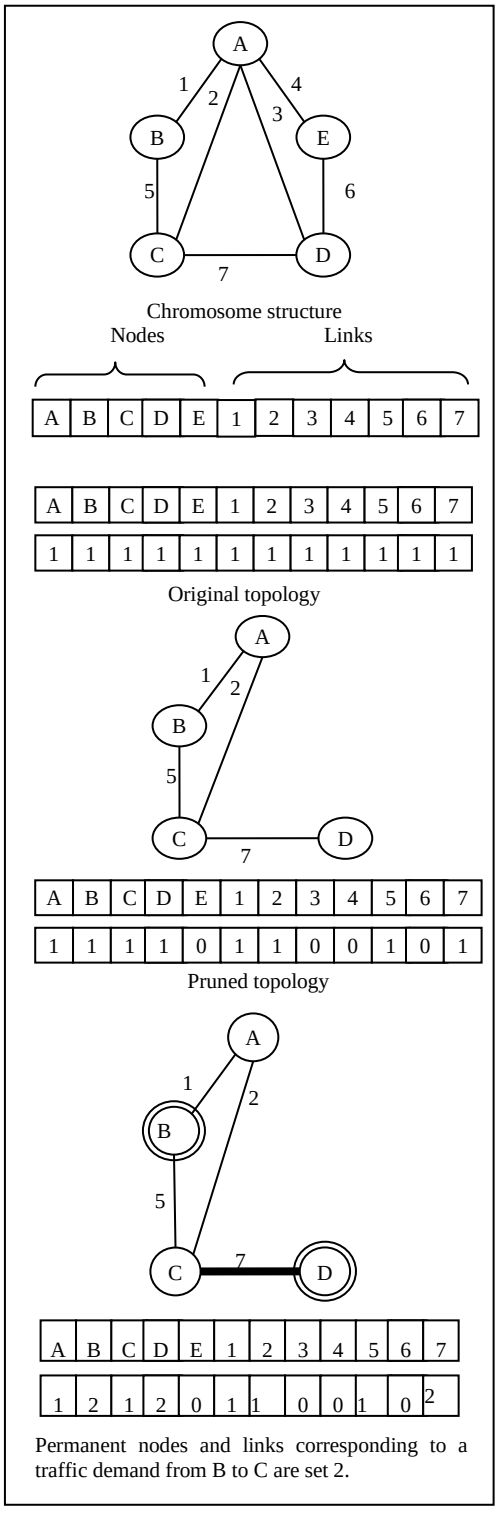


Fig. 3. GA representation of a network

A reproduction operation is a function that gets an individual or two individuals as input and outputs a set of new individuals. Usually, two types of reproduction operations are used. A *mutation* function alters a single gene of an individual. Mutation introduces more diversity into the population. *Crossover* function exchanges portions of two individuals and generates their children in the hope the children are more feasible than their parents. We use two-point cross-over and single-point mutation, as depicted in Figure 4.

### 3.4.5 Stopping Criteria

Selection and reproduction operations are repeated several times over the current generation to induce a new generation. Each subsequent generation is built with the hope that its members would be better than the previous generation.

The process of evolution through generations stops when some set of sufficiently good individuals are found or there is a little or no hope to find suitable individuals in the future.

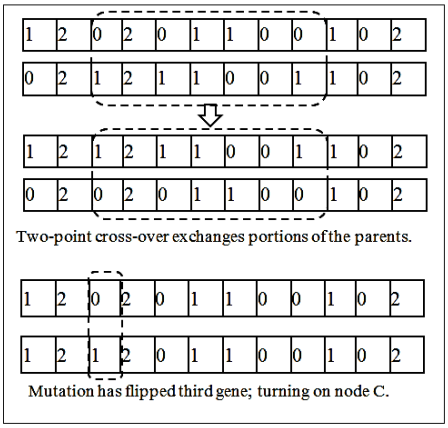


Fig. 4. GA evolutionary operations

## 4. Simulation and Performance Evaluation

Proposed algorithms are implemented in C-language, encapsulating TOTEM toolbox's DAMOTE algorithm[1] [7]. TOTEM is an open source toolbox for modeling, simulation, and evaluation of protocols and algorithms for network traffic engineering. DAMOTE is a traffic engineering package featuring a configurable score function and rich report generation. Among various quantitative and qualitative attributes of topologies, statistics such as *maximum links utilization*, *total sum of network energy consumption*, and number *of inactive links and nodes* are of particular interest in this research.

For genetic algorithms, we used populations with 1000 members evolving through 20 generations. Cross-over rate was 1, since we were to generate a population of predefined size, and mutation rate was 0.01.

### 4.1 Evaluation Information

Abilene topology and its traffic matrices were used in experiments [26]. Since all nodes in this topology are edge nodes (i.e. flow sources or destinations), to test deactivation feature of the proposed algorithms an augmented topology is generated by adding virtual core nodes to the original Abilene topology. The new topology is called extended Abilene. Table 1 summarizes specifications of Abilene and Extended Abilene networks.

1. http://totem.run.montefiore.ulg.ac.be/algos/damote.html

Table 1. Specifications of Abilene and Extended Abilene Networks

| Name | #Nodes | #Links | #Edge Nodes | #Core Nodes | %Core Nodes |
|---|---|---|---|---|---|
| Abilene | 12 | 30 | 12 | 0 | 0 |
| Extended Abilene | 21 | 52 | 12 | 9 | 43 |

Following the idea of [16], traffic matrices of these networks are scaled by 10 so that maximum link utilization reaches up to 50%. Scaling traffic matrix makes the proposed algorithms more robust to variations in traffic patterns. Figure 5 shows the utilization of links in extended Abilene network at 22:00 on December 15, 2007. It confirms that most link utilizations are below 0.5. For the same network with 10-times scaled traffic matrix, maximum network utilization in different hours of a day is displayed in Figure 6. The figure confirms that maximum network utilization is below 0.6 most of the time. Over-plotted dashed line is the trend line of the maximum network utilization. Its slightly increasing trend confirms that network traffic is not stationary; hence there is no single best network topology for all hours of a day.

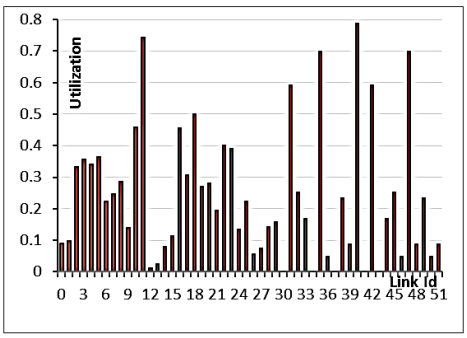


Fig. 5. Link utilization in extended Abilene

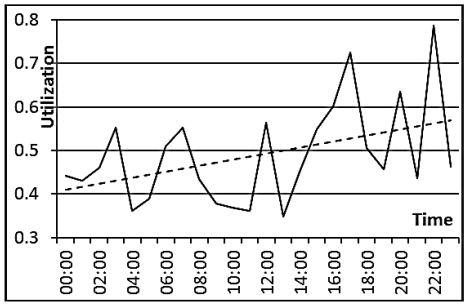


Fig. 6. Actual hourly network utilization for extended Abilene

## 4.2 Network Links Utilization

The four proposed algorithms are applied to the extended Abilene network and results are reported hereafter. Maximum network utilization during a day is displayed in Figure 7. For clarity of the demonstration, their respective trend lines are plotted in Figure 8. EMST has higher utilization than other algorithms. After that, SLE has a similar trend. EAD has utilization around 0.68 with a constant trend that makes it a stable algorithm. Among all, EAGA has the steepest trend denoting its adaptive behavior against traffic variations.

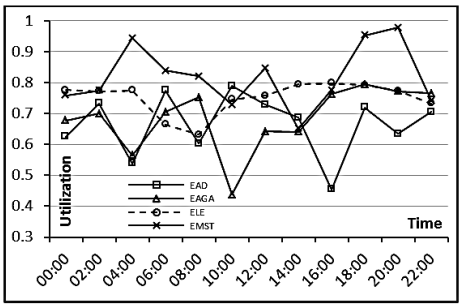


Fig. 7. Hourly network utilization

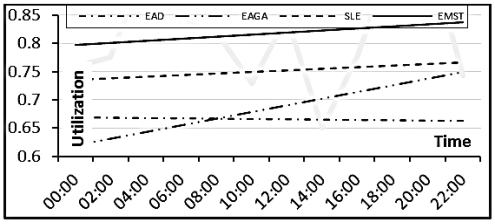


Fig. 8. Trends of network utilization (see fig. 7)

## 4.3 Power Saving

The amount of power saving of the four algorithms is presented in Figure 9 and Table 2. Figure 10 and Figure 11 report percentage of deactivated links and nodes for various algorithms, respectively.

As before, ELE, EAGA, and EMST have similar patterns in power saving. They achieve almost 38% power saving in most of the hours in a day. EAD has lower saving rate as compared to others. This may be due to the fact that it tries to keep maximum link utilization as low as possible. This requires incorporating more links and nodes, hence using more energy.

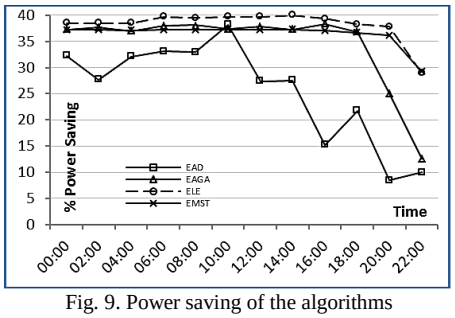


Fig. 9. Power saving of the algorithms

Table 2. Number of deactivated links and nodes

| | EAGA | | EMST | | ELE | | EAD | |
|---|---|---|---|---|---|---|---|---|
| | N_off | L_off | N_off | L_off | N_off | L_off | N_off | L_off |
| 00:00 | 6 | 28 | 6 | 28 | 6 | 30 | 5 | 26 |
| 02:00 | 6 | 29 | 6 | 28 | 6 | 30 | 4 | 26 |
| 04:00 | 6 | 27 | 6 | 28 | 6 | 30 | 5 | 26 |
| 06:00 | 6 | 30 | 6 | 28 | 6 | 33 | 5 | 29 |
| 08:00 | 6 | 30 | 6 | 28 | 6 | 33 | 5 | 28 |
| 10:00 | 6 | 28 | 6 | 28 | 6 | 33 | 6 | 31 |
| 12:00 | 6 | 29 | 6 | 28 | 6 | 33 | 4 | 25 |
| 14:00 | 6 | 28 | 6 | 28 | 6 | 34 | 4 | 25 |
| 16:00 | 6 | 31 | 6 | 28 | 4 | 30 | 2 | 16 |
| 18:00 | 6 | 27 | 6 | 27 | 6 | 30 | 3 | 22 |
| 20:00 | 4 | 20 | 6 | 27 | 2 | 21 | 1 | 10 |
| 22:00 | 2 | 10 | 5 | 18 | 3 | 22 | 1 | 14 |

N_off: #deactivated nodes, L_off: #deactivated links

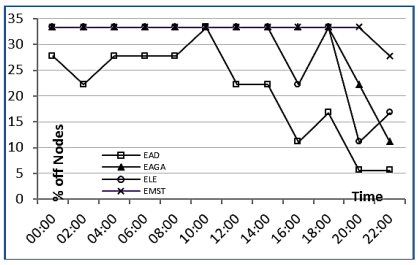


Fig. 10. Percentage of inactive nodes

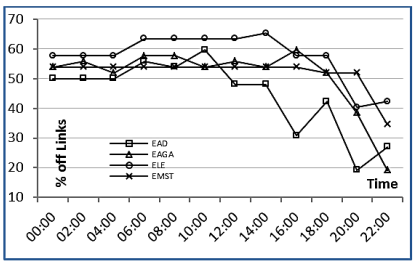


Fig. 11. Percentage of inactive links in extended Abilene

## 4.4 Comparison with Related Works

We compared the performance of our proposed algorithms against that of two notable recent works, one based on OSPF (Amaldi et al. [16]) and the other one based on MPLS (Ruiz-Rivera et al. [17]).

As mentioned earlier, Amaldi et al. alter link weights in order to direct the routing procedure to find more energy efficient paths. They report their experiments with two traffic matrices over Abilene and extended Abilene networks. One for 6:00 that denotes a low traffic load and the other for 12:00 that represents the highest load.

Table 3. Comparison of our algorithms against Amaldi et al. [16]

| | 06:00 AM (low traffic) | | | | 12:00 AM (high traffic) | | | |
|---|---|---|---|---|---|---|---|---|
| | L-off | N-off | %Saving | %MLU | L-off | N-off | %Saving | %MLU |
| ELE | 33 | 6 | 37 | 66 | 30 | 6 | 37 | 78 |
| EMST | 28 | 6 | 37 | 84 | 28 | 6 | 37 | 76 |
| EAGA | 28 | 6 | 38 | 70 | 30 | 6 | 37 | 68 |
| EADA | 29 | 5 | 33 | 77 | 26 | 5 | 32 | 62 |
| Amaldi | 54 | 7 | 54 | 42 | 40 | 3 | 21 | 78 |

Table 3 summarizes results for the compared algorithms. Amaldi et al. works well for low traffic scenarios and turns off much more links as compared to ours. However number of turned off nodes are similar. For heavily loaded scenarios, our algorithms work much better than Amaldi's.

Ruiz-Rivera et al. enhance energy consumption in MPLS networks by favoring LSPs that share links with previous or future requests. They report link utilization for various levels of bound width requests. As the required bandwidth increases, maximum link utilization increases as well. Since their experimental setup and results format is significantly different from ours (and also from Amaldi et al.) a comprehensive comparison and discussion of the results is not applicable. However, we compare the rate at which link utilization increases in response to the traffic increments. Table 4 reports mean and standard deviation of maximum link utilization for studied methods. For lower traffics, link utilization is low. Smaller mean values denote that the method keeps MLU at lower levels by incorporating more links. This results in more energy consumption in the network. Hence, lower MLU roughly translates to lower energy saving. With this observation, ELE and Ruiz-Rivera seem to be more energy efficient than others.

Focusing on MLU deviations, lower deviation means that the algorithm uses links in accordance with requested traffic load. That is, if the traffic increases, more links are employed to keep MLU within bounds. However, higher deviation means that the set of active links are decided in advance and their utilization varies in response to request variations. With this observation, it can be concluded that Ruiz-Rivera uses many lightly loaded links for low-traffic scenarios. That is, the approach saves more energy for heavy traffic scenarios but is not energy efficient for lightly loaded cases.

Table 4. Maximum link utilization for compared algorithms

| | ELE | EMST | EAGA | EADA | Amaldi | Ruiz |
|---|---|---|---|---|---|---|
| Mean | 78 | 82 | 68 | 67 | 65 | 79 |
| Std. | 5 | 10 | 10 | 10 | 15 | 24 |

## 4.5 Adaptive Topology Planning

The ultimate goal of developing these algorithms is planning and proposing a topology for a network that is more energy-efficient than the original network. The proposed topology must be robust to traffic fluctuations and save energy as much as possible. The central idea in planning is to define a profile as a set of conditions for the network. Then, a topology is constructed and archived for each network profile. This set is built offline using the proposed algorithms. Being offline allows us to employ sophisticated simulation and estimation techniques without much worrying about the time complexities.

As a simple implementation of this approach, we consider temporal profiles. As mentioned earlier, traffic matrices used in these experiments are for 12 two-hour intervals, namely 00 to 02, 02 to 04, ..., 22 to 00. Using any of the proposed algorithms the best topology is constructed for each interval. This topology has a set of links and nodes that are subject to deactivation. To make smooth changes to the network, and hence make it more stable, a day is partitioned into three equal intervals, namely 0 to 8, 8 to 16, and 16 to 0. Each of these eight-hour intervals spans four two-hour intervals for which there exists a special topology. Then, a topology is proposed for eight-hour intervals that excludes inactive nodes and links of all the corresponding four topologies.

Profile-based planning enables dynamic selection of proper topology for current network status. If a traffic demand cannot be handled by the current topology, another topology with more bandwidth resources will be enabled in which more links and/or nodes are active. The new profile can be switched proactively by estimation of network status in the near future. For example, if current maximum link utilization goes beyond 90% a profile with more resources can be used to prevent possible future failures.

## 5. Conclusion and Future Directions

In the scope of green networking, this paper proposed four algorithms for adaptive and dynamic selection of topologies for backbone data networks. These algorithms try to tailor down the network so that besides serving demanded traffic, a set of its nodes and links are deactivated to save energy.

The proposed algorithms demonstrated feasibility and usefulness of automatic topology adaptation. Sequential Link Elimination (SLE), Extended Minimum Spanning Tree (EMST), and Energy-Aware DAMOTE (EAD) are deterministic algorithms that serially process nodes and links for removal. SLE works in backward manner in which at first all the links and nodes are engaged. In subsequent steps, it tries to turn off links while preserving network status in serving all the demanded traffic. EMST takes a reverse approach by constructing minimum spanning tree of the network. Initially, all the links out of this tree are inactive. The algorithm adds links to the tree until it can handle the demanded traffic. EAD is similar to SLE in the sense that it sequentially processes the links. The difference is that in EAD, relative order of links is determined by both their energy saving and maximum utilization features.

SLE, EMST, and EAD are greedy algorithms. Greedy algorithms usually find a suboptimal solution that can be very different from optimal solutions. On the other hand, Genetic Algorithms (GAs) randomly search solution space for proper suboptimal solutions. Usually, a GA with a sufficiently large population size and generations finds near-optimal solutions. However, the computational cost of running GAs is more than the deterministic algorithms.

Experiments on extended Abilene network confirmed that the proposed algorithms are capable of saving a considerable amount of energy consumption in the network. Analysis revealed that EAD is best in controlling congestion and EAGA is better in saving more energy.

In the future, current research can be extended in several dimensions. At first, a distributed and online solution in which each core node reactively decides by itself to enable or disable its outgoing links is desirable. The node inspects the current network status and performs an action to make it better. Reinforcement learning based approaches can model this behavior very efficiently. However, it would require plenty of training network history to make good decisions.

This paper has focused on the net amount of energy that is used by the network. Consider a network that some of its elements use solar energy and some others use fossil-fuel electricity. An extension of current research would be taking to account natural preferences of these sources (e.g. solar to fossils).

Genetic algorithms are computation intensive by nature. Recently, parallel implementation of GAs on multiprocessor systems or special purpose boards such ad graphical processing units (GPUs) has caught a considerable interest. Online implementation of GAs on newer technologies and modeling complex and feature-rich networks by GAs is another direction for future research.

## References


[1] J. Baliga, K. Hinton and R. S. Tucker, Energy consumption of the Internet, University of Melbourne, Department of Electrical and Electronic Engineering, 2011.

[2] C. Fang, F. R. Yu, T. Huang, J. Liu and Y. Liu, "A Survey of Green Information-Centric Networking: Research Issues and Challenges.," IEEE Communications Surveys and Tutorials, vol. 17, pp. 1455-1472, 2015.

[3] I. F. Akyildiz, A. Lee, P. Wang, M. Luo and W. Chou, "A roadmap for traffic engineering in SDN-OpenFlow networks," Computer Networks, vol. 71, pp. 1-30, 2014.

[4] A. G. Cervero, M. Chincoli, L. Dittmann, A. Fischer, A. E. Garcia, J. Gal{\'a}n-Jim{\'e}nez, L. Lefevre, H. d. Meer, T. Monteil, P. Monti and others, "Green wired networks," Large-Scale Distributed Systems and Energy Efficiency: A holistic view, pp. 41-80, 2015.

[5] M. Baldi and Y. Ofek, "Time for a" greener" internet," in Communications Workshops, 2009. ICC Workshops 2009. IEEE International Conference on, 2009.

[6] R. Bolla, R. Bruschi, F. Davoli and F. Cucchietti, "Energy efficiency in the future internet: a survey of existing approaches and trends in energy-aware fixed network infrastructures," Communications Surveys \& Tutorials, IEEE, vol. 13, pp. 223-244, 2011.

[7] G. Leduc, H. Abrahamsson, S. Balon, S. Bessler, M. D'Arienzo, O. Delcourt, J. Domingo-Pascual, S. Cerav-Erbas, I. Gojmerac, X. Masip and others, "An open source traffic engineering toolbox," Computer Communications, vol. 29, pp. 593-610, 2006.

[8] F. Blanchy, L. M{\'e}lon and G. Leduc, "A Preemption-aware on-line routing algorithm for MPLS networks," Telecommunication Systems, vol. 24, pp. 187-206, 2003.

[9] P. Mahadevan, P. Sharma, S. Banerjee and P. Ranganathan, "A power benchmarking framework for network devices," in NETWORKING 2009, Springer, 2009, pp. 795-808.

[10] K. Cengiz and T. Dag, "A review on the recent energy-efficient approaches for the Internet protocol stack," EURASIP Journal on Wireless Communications and Networking, vol. 2015, pp. 1-22, 2015.

[11] A. P. Bianzino, C. Chaudet, D. Rossi and J. L. Rougier, "A Survey of Green Networking Research," IEEE Communications Surveys Tutorials, vol. 14, pp. 3-20, 2012.

[12] M. Xia, M. Tornatore, Y. Zhang, P. Chowdhury, C. U. Martel and B. Mukherjee, "Greening the Optical Backbone Network: A Traffic Engineering Approach.," in ICC, 2010.

[13] G. Lin, S. Soh, M. Lazarescu and K.-W. Chin, "Energy-aware two link-disjoint paths routing," in High Performance Switching and Routing (HPSR), 2013 IEEE 14th International Conference on, 2013.

[14] B. Addis, A. Capone, G. Carello, L. G. Gianoli and B. Sanso, "Energy-aware multiperiod traffic engineering with flow-based routing," in Communications (ICC), 2012 IEEE International Conference on, 2012.

[15] W. Fisher, M. Suchara and J. Rexford, "Greening backbone networks: reducing energy consumption by shutting off

cables in bundled links," in Proceedings of the first ACM SIGCOMM workshop on Green networking, 2010.
[16] E. Amaldi, A. Capone and L. G. Gianoli, "Energy-aware IP traffic engineering with shortest path routing," Computer Networks, vol. 57, pp. 1503-1517, 2013.
[17] A. Ruiz-Rivera, K.-W. Chin, S. Soh and R. Raad, "On the performance of online and offline green path establishment techniques," EURASIP Journal on Wireless Communications and Networking, vol. 2015, pp. 1-17, 2015.
[18] B. Fortz, "Applications of meta-heuristics to traffic engineering in IP networks," International transactions in operational research, vol. 18, pp. 131-147, 2011.
[19] E. Bonetto, L. Chiaraviglio, F. Idzikowski and E. Le Rouzic, "Algorithms for the multi-period power-aware logical topology design with reconfiguration costs," Journal of Optical Communications and Networking, vol. 5, pp. 394-410, 2013.
[20] R. Samadi, M. Nassiri and M. Mansoorizadeh, "Energy-aware traffic engineering in IP networks using non-dominated sorting genetic II algorithm," International Journal of Advanced Intelligence Paradigms, vol. 1, 2019.
[21] R. Hassan, B. Cohanim, O. De Weck and G. Venter, "A comparison of particle swarm optimization and the genetic algorithm," in 46th AIAA/ ASME/ ASCE/ AHS/ ASC structures, structural dynamics and materials conference, 2005.
[22] O. Okonor, N. Wang, S. Georgoulas and Z. Sun, "Green Link Weights for Disruption-Free Energy-Aware Traffic Engineering," IEEE Systems Journal, vol. 11, pp. 661-672, 2017.
[23] R. Wang, S. Gao, W. Yang and Z. Jiang, "Energy aware routing with link disjoint backup paths," Computer Networks, vol. 115, pp. 42-53, 2017.
[24] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, Introduction to algorithms third edition, MIT Press, 2009.
[25] A. P. Engelbrecht, Computational intelligence: an introduction, John Wiley & Sons, 2007.
[26] S. Orlowski, M. Pi{\'o}ro, A. Tomaszewski and R. Wess{\"a}ly, "SNDlib 1.0--Survivable Network Design Library," in Proceedings of the 3rd International Network Optimization Conference (INOC 2007), Spa, Belgium, 2007.

**Mojtaba Sabahi-Aziz** Got his BSc from Shamsipour faculty of engineering in software engineering in 2013. He earned his MSc in computer networks from Bu-Ali Sina University in 2016. His research interests include monitoring and evaluation of computer networks.

**Sepideh Zarei** Got her BSc in software engineering in 2013 and MSc in computer networks in 2016 both from Bu-Ali Sina University. Her research interests include monitoring and evaluation of computer networks.

**Muharram Mansoorizadeh** is an assistant professor at the Computer Department of Bu-Ali Sina University. He received his BSc degree in software engineering from the University of Isfahan, Isfahan, Iran, in 2001, and his MSc degree in software engineering and the PhD in computer engineering from Tarbiat Modares University, Tehran, Iran, in 2004 and 2010, respectively. His current research interests include machine learning, affective computing and information retrieval.

**Mohammad Nassiri** is an assistant professor at the Computer Department of Bu-Ali Sina University. He received his BSc and MSc degrees from Iran University of Science and Technology (IUST) and Sharif University of Technology, respectively, in 2000 and 2002. He also received his Ph.D. in Computer Engineering from Grenoble INP, France, in 2008. His current research interests include Performance Evaluation of wireless networks and the Internet of things, Internet Traffic Classification and Software networking paradigm.

# Lifetime Improvement Using Cluster Head Selection and Base Station Localization in Wireless Sensor Networks

Maryam Najimi*
Department of Electrical and Computer Engineering of University of Science and Technology of Mazandaran (USTM), Behshahr, Iran
Maryam_najimi1361@yahoo.com
Sajjad Nankhoshki
Department of Electrical and Computer Engineering of University of Science and Technology of Mazandaran (USTM), Behshahr, Iran
sajjadnankhoshki@yahoo.com

Received: 13/Mar/2018 Revised: 01/Sep/2018 Accepted: 31/Oct/2018

**Abstract**

The limited energy supply of wireless sensor networks poses a great challenge for the deployment of wireless sensor nodes. In this paper, a sensor network of nodes with wireless transceiver capabilities and limited energy is considered. Clustering is one of the most efficient techniques to save more energy in these networks. Therefore, the proper selection of the cluster heads plays important role to save the energy of sensor nodes for data transmission in the network. In this paper, we propose an energy efficient data transmission by determining the proper cluster heads in wireless sensor networks. We also obtain the optimal location of the base station according to the cluster heads to prolong the network lifetime. An efficient method is considered based on particle swarm algorithm (PSO) which is a nature inspired swarm intelligence based algorithm, modelled after observing the choreography of a flock of birds, to solve a sensor network optimization problem. In the proposed energy- efficient algorithm, cluster heads distance from the base station and their residual energy of the sensors nodes are important parameters for cluster head selection and base station localization. The simulation results show that our proposed algorithm improves the network lifetime and also more alive sensors are remained in the wireless network compared to the baseline algorithms in different situations.

**Keywords:** Wireless Sensor Nodes; Network Lifetime; Particle Swarm Algorithm (PSO); Base Station; Cluster Head.

## 1. Introduction

Nowadays, wireless sensor networks (WSNs) are used in various applications such as military tracking, environmental monitoring, medical diagnosis and habitual monitoring, etc. The significant role of the sensor nodes in these networks include data gathering from the environment and send it to the base station (BS) to make a final decision about the status of the environment. However, energy consumption of the sensor nodes is one of the main concerns in wireless sensor networks. Clustering is one of the most efficient techniques to save energy in these networks. In this technique, sensor nodes transmit their data to some leader nodes called cluster heads (CHs). CHs send aggregated data to BS in one-hop communication. However, in the process of clustering, selection of the CHs performs a very crucial role for saving the energy and therefore improving the network lifetime as it has several impacts on the energy conservation of member sensor nodes. Several number of clustering algorithms based on heuristic methods have been developed for WSNs [1-4]. LEACH algorithm is one of the well-known distributed clustering algorithm, in which a cluster head is selected with some probability among the sensor nodes to save more energy, however, the remaining energy is not considered for cluster head selection. Therefore, it is possible to select a sensor node with low energy as a cluster head which leads to die quickly [1]. In [5], Centralized LEACH is proposed which the distance and the energy of the nodes are also considered in cluster head selection. In [6], the cluster head selection is done using particle swarm algorithm (PSO) algorithm and the ratio of total initial energy of all nodes to the total current energy of the all cluster heads is considered, however, the distance from sink is not considered. In [7], residual energy, distance and node density is considered in cluster head selection. However, the cluster formation phase is ignored which it leads to have high energy consumption. In [8], a novel Energy Efficient Connected Coverage (EECC) scheduling is proposed to maximize the lifetime of the WSN. The EECC adheres to Quality of Service (QoS) metrics such as remaining energy, coverage and connectivity. In EECC the sensor which doesn't contribute to coverage will act as a relay node to reduce the burden of the sensing node. In [9], the paper introduces an algorithm named Fuzzy logic based unequal clustering, and Ant Colony Optimization (ACO) based Routing, Hybrid protocol for WSN to eliminate hot spot problem and extend the network lifetime. This protocol comprises of Cluster Head (CH) selection, inter-cluster routing and cluster maintenance. In [10], Extended-Multilayer Cluster Designing Algorithm (E-MCDA) approach is proposed in a large network. Performance of E-MCDA is evaluated in energy consumption at various aspects of energy, packets transmission, the number of designed clusters, the number of nodes per cluster and un-clustered nodes.

* Corresponding Author

Another important issue is the base station (BS) localization which is a critical factor in designing a wireless sensor network. Using this method, less power consumes to deliver the cluster head's data to BS. As a consequence, the network lifetime is improved. However, in some papers, BS location is deployed within the center point of the area of interest [11]. Although it gives fast suitable solutions, it cannot guarantee the optimal BS location. In [12], two algorithms are proposed to determine the optimal location of the base station; for homogeneous nodes as well as for heterogeneous nodes where the single-hop routing is used. The limitation of this approach is the high energy consumptions of the nodes located far from the sink. In [13], one approach is proposed based on PSO algorithm for determining the best position of the sink where multi-hop communication is considered. But in multi-hop communication the data collected by all the sensors reach the sink through the nodes close to the sink and thus these nodes may die soon due to pass a huge amount of data.

Therefore, our contribution in this paper is as follows:

- At first, we propose an energy efficient BS localization using PSO algorithm. Then, for saving more energy, the problem of the cluster heads selection is considered. In this case, sensors send their data to their corresponding cluster heads. The number of cluster heads which are selected among the sensors, are fixed. However, for cluster head selection, the remaining energy of each sensor and also its distance from BS is considered.
- After the cluster head selection, the proper position of BS is also determined using PSO algorithm to conserve more energy. Simulation results show the effectiveness of the proposed algorithm in improving the network lifetime.

The remainder of this paper is organized as follows. The network model is detailed in Section 2. The overview of PSO (Particle Swarm Optimization) technique is stated in section 3. The proposed algorithms based on PSO algorithm are shown in section 4. Performance evaluation results that demonstrate the efficiency of the proposed algorithms are presented in Section 5. Conclusions are drawn in Section 6.

## 2. Network Model

We consider a WSN deployed in a square area with a set of normal sensor nodes and high energy Base Station (BS). Normal sensor nodes sense local data about the environment and forward them to BS. Nodes can be deployed manually or randomly in the target area (Fig.1).

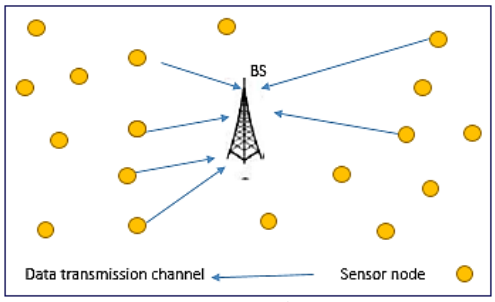


Fig. 1. Sensor node locations for data transmission.

As we said, sensor nodes in wireless sensor networks are energy constrained and cannot be rechargeable. Since battery is the only power source to the sensors, their energy should be carefully utilized to increase the network lifetime and improve its performance. The energy model used in this paper is based on two parameters [1]: $E_{t-elec}$ is the transmitter electronics energy, $E_{amp}$ is the required amplification. The energy consumption of each node depends on the amount of the data and also distance between each node and its receiver. In this paper, we assume that $L$ reliable bits are transmitted to BS. Therefore, the total energy consumption is obtained as follows [14].

$$E_T = \sum_{j=1}^{N} L(E_{t-elec} + E_{amp} {d_j}^2) \quad (1)$$

Where $N$ is the number of sensors and $d_j$ is the distance between the $j$ th node and BS. According to Eq.(1), the energy consumption is related to the distance between each node and BS. On other hand, BS should be located in a proper position from the sensors to save more energy. In fact, the suitable location of BS leads to increase the residual energy of the nodes. Therefore, the network lifetime is improved. We define the network lifetime to be the time until 25 percent of the sensors run out of energy [15], [16]. For this purpose, we use PSO algorithm as a random optimization algorithm to find the optimal position of BS. This algorithm was developed through the inspiration of social behavior of birds flocking. The details of this algorithm is at the next section.

## 3. Overview of PSO

PSO is a nature inspired swarm intelligence based algorithm, modelled after observing the choreography of a flock of birds, i.e., how they can explore and exploit the multi-dimensional search space for food and shelter [17], [18]. PSO consists of a predefined number of particles, $M$, called swarm. This algorithm searches the area with respect to the mathematical formula over velocity and position of each particle, $P_i, 1 \leq i \leq M$. A fitness function is used to evaluate each particle for verifying the quality of the solution. The objective of PSO is to find the particle's positions that result best evaluation of the given fitness function. PSO is initialized with a group of random particles (solutions) and then searches for optimal solution by updating iterations. In each iteration, each particle finds its own best, i.e., personal best called $Pbest_i$. Another "best" value that is tracked by the particle swarm optimizer is the best value, obtained so far by any particle in the population. This best value is a global best and called $Gbest$. In each iteration, velocity of each particle is updated using the current velocity of the particle and the previous local best and global best position. Depending upon the past value, new velocity and new position of the particles can be estimated. The same procedure is repeated for each iteration. The

formula for updating velocity and position of each particle is given by the following equations [19].

$$V_{p_i}(k+1) = w.V_{p_i}(k) + c_1.r_1(k)\left(Pbest_i - X_{p_i}(k)\right) + c_2.r_2(k)\left(Gbest - X_{p_i}(k)\right) \quad (2)$$

And

$$X_{p_i}(k+1) = V_{p_i}(k+1) + X_{p_i}(k) \quad (3)$$

Where, $V_{p_i}$ and $X_{p_i}$ are the new velocity and position of the $i$th particle, respectively. $w$ is the inertia weight, $c_1$ and $c_2$ are the acceleration coefficients and $r_1(k)$ and $r_2(k)$ are random numbers uniformly distributed in [0,1]. The updating process is repeated until it is reached to an acceptable value of $Gbest$. After getting new updated position, the particle evaluates the fitness function and updates $Pbest_i$ as well as $Gbest$ for the minimization problem as follows [20].

$$Pbest_i = \begin{cases} E_{T_i} & \text{if fitness}(P_i) < Pbest_i \\ Pbest_i \text{ doesnot change} & \text{otherwise} \end{cases} \quad (4)$$

$$Gbest = \begin{cases} Pbest_i & \text{if } Pbest_i < Gbest \\ Gbest \text{ doesnot change} & \text{otherwise} \end{cases} \quad (5)$$

# 4. Proposed Algorithm

## 4.1 Proposed Algorithm without Clustering

As we said, in this paper, our aim is improving the network lifetime. For this purpose, we use PSO algorithm to find the optimal location of BS. According to PSO algorithm, each particle, $P_i$, shows the random coordinate of BS which lies in the corresponding environment. According to this position, the energy consumption of each sensor node for data transmission to each particle is calculated. The fitness function is the total energy consumption of all sensors in data transmission to BS. Our aim is to minimize the fitness function by determining the best location of BS. Therefore, the personal best (i.e., $Pbest_i$) is calculated for each particle through their fitness value. Then, the global best (i.e., $Gbest$ ) is calculated based on the $Pbest_i$ values according to Eq.(5). In learning algorithm, in each iteration the velocity and position of each particle are updated according to Eq.(2) and Eq.(3), respectively. Then, the fitness function is calculated again according to Eq.(1). According to this function, $Pbest_i$ and $Gbest$ are obtained. It should be noted that the sensor nodes are participated in data transmission which have enough energy. It means that their remaining energy is more than their energy consumption. The algorithm ends when the termination criteria is fulfilled. The pseudo code for the PSO-Based BS Localization Algorithm (PBBSL) is as below.

**PBBSL Algorithm**

```
While (Number of alive nodes>0.25*total number of nodes)
  Determine the nodes which their remaining energy are more than
their  energy consumption
  Initialize the particles P_i ∀i ∈ M(M is the number of particles)
  For i = 1:M
  Calculate the Fitness (E_Ti)for each P_i
   Pbest_i = E_Ti
  End
  Gbest =min(E_Ti)
  While(termination criteria is not satisfied)
    For i = 1:N
  Update velocity and position of the particles P_i
  Calculate Fitness(E_Ti)
  If Fitness of P_i< Pbest_i
       Pbest_i=E_Ti
  If Fitness of P_i< Gbest
      Gbest= Pbest_i
   End
   End`
    End
  End
  Compute the energy consumption for each node
  Compute the remaining energy for each sensor
  End while
```

Fig. 2. Pseudo code for PBBSL algorithm

## 4.2 Proposed Algorithm with Clustering

In our network model, it is possible that some sensing nodes are located in far distances from BS. Therefore, energy consumption for transmitting data to BS increases. In our system model one solution for saving energy is that, sensor nodes send their data to the cluster heads (CH). CH is a node among all sensors and it sends the data of the sensor nodes to BS.

In this section, the main purpose is selecting the cluster heads by considering the energy efficiency so that the network lifetime is improved. For cluster head selection, the residual energy of the sensor nodes and also the distance of each cluster head with other nodes which transmit data to the cluster head, are considered. Then, according to the cluster heads' position, the best location of BS is obtained according to PSO algorithm. For cluster head selection, the environment is divided at most into four squares and four nodes with enough energy and nearest to the middle of the squares are candidates as CHs. It should be noted that the environment is divided according to the number of the alive nodes. It means that for the empty square, there is not any cluster head. After the cluster head selection, finding the best location of BS is the similar to PBBSL algorithm, except that, the BS location is obtained according to the CHs' position. The flowchart of the proposed algorithm is shown in Fig. 3.

# 5. Performance Evaluation

In our wireless network, nodes are uniformly distributed in a square field with a length of 200 m. $E_{int} = 0.2mJ$ is assumed as the initial energy for each sensor and the alive nodes are considered as the nodes that their remaining energies are more than the energy required for data transmission to CHs. CHs send the

results of data transmitted from their corresponding nodes to BS. $L = 10$ bits are considered for data transmission. In all comparisons the exact optimal results are numerically obtained in MATLAB. Every simulation result in this section is averaged over 10000 realizations. We model the wireless channel between each sensor and CHs and also between each CH and BS using a free-space path loss model. By assuming a data rate of 250 kb/s and a transmit power of 20 mW, we consider $E_{t-elec}$=80 nJ [21],[22]. The $E_{amp}$ to satisfy a receiver sensitivity of -90 dBm is $40.4 pJ/m^2$ [23].The parameters of PSO are in table1.

Table 1. Parameters for PSO in simulation

| Parameter | Value |
|---|---|
| $c_1$ | 2 |
| $c_2$ | 2 |
| w | 0.3 |
| M | 10 |

For showing the effectiveness of our proposed algorithms, we compare them with the following algorithms:

- Random BS Localization (RBSL) Algorithm: In this algorithm, BS position is selected randomly in each iteration. Then, the sensors send their data to BS. This algorithm has the minimum complexity to find the solution for our problem.
- Fixed BS Localization (FBSL) Algorithm: In this algorithm, BS location is fixed in environment and its position does not change according to the alive nodes location. This algorithm is considered to show that determining the location of BS improves the network lifetime.

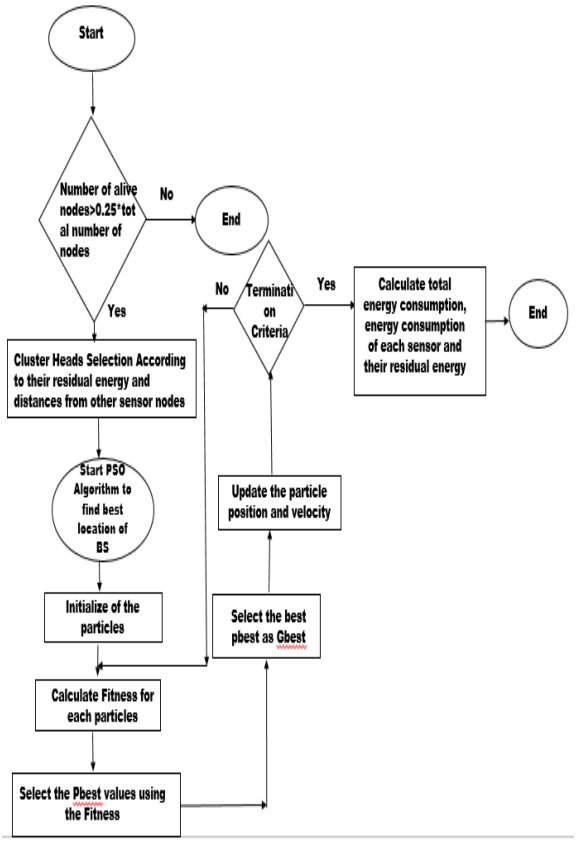


Fig. 3. Flowchart of the proposed algorithm with CHs selection

Fig. 4 shows the number of alive nodes for different algorithms. It is clear that in PBBSL algorithm with clustering, number of alive nodes is more than the other algorithms. PBBSL algorithm also have more alive sensors than RBSL and FBSL algorithms due to the random and fixed position of BS. In this case, it is possible to locate more sensors far from BS and therefore, the energy consumption is increased significantly. In fact, more alive sensors states more lifetime for the network. On the other hand, cluster heads and BS localization lead to improve the network lifetime. The dimension of the environment is set to 1000m.

Fig. 5 shows the average total remaining energy of the nodes. It is clear that cluster head selection leads to save more energy in sensors due to the decreasing the distance for data transmission. We also note that determining the location of BS according to the CHs position helps to have more remaining energy for CHs. It should be noted that in less number of nodes, cluster head selection is not effective in saving energy, however, increasing the number of sensors show the effectiveness of the cluster head selection in having more remaining energy. According to the results, RBSL and FBSL algorithms have lower remaining energy. On the other hand, they have lower alive sensors due to the random position and fixed position for BS, respectively.

Fig. 6 shows the average total energy consumption versus different nodes. According to the results, CHs selection and determining the location of BS increases the energy consumption of the nodes for data transmission to CHs and also data transmission from CHs to BS. It shows that there are more alive sensors to transmit their data to the cluster heads or BS while in FBSL and RBSL, the energy consumption is decreased because less nodes are still alive to send their data to their destination. It should be noted that as the number of the sensors increases, more energy consumes due to the existence of more alive nodes in the environment.

In Fig. 7 the total remaining energy of the sensors for different algorithms is shown. In fact, this metric states the successful percent of the algorithms in balancing the energy consumption of the alive sensors. Our proposed algorithms have the highest value of the total remaining energy. Therefore, the network lifetime is improved. According to the results, by increasing the dimension of the environment, the total remaining energy is increased. Because, it is possible to have more sensors far from BS and hence, the energy consumption increases. The number of sensors is set to 50.

Fig. 8 shows the number of alive nodes for different algorithms. In fact, this metric states the role of the algorithms for increasing the network lifetime. According to the results, our proposed algorithms have the most network lifetime while FBSL and RBSL algorithms have the least value. It means that clustering and BS localization improve the network lifetime significantly. It should be noted that only the sensors with enough energy participate in data transmission.

Fig. 9 shows the total energy consumption of different algorithms versus different environments. Although, our proposed algorithms consume more energy for data transmission, however, there is a balance between the sensors for sending data. On the other hand, more alive nodes consume more energy to transmit their data to BS or cluster heads. FBSL and RBSL algorithms consume less energy due to the less number of alive nodes. It should be noted that as the dimension of the environments increases the energy consumption increases due to the more distances between sensors and cluster heads.

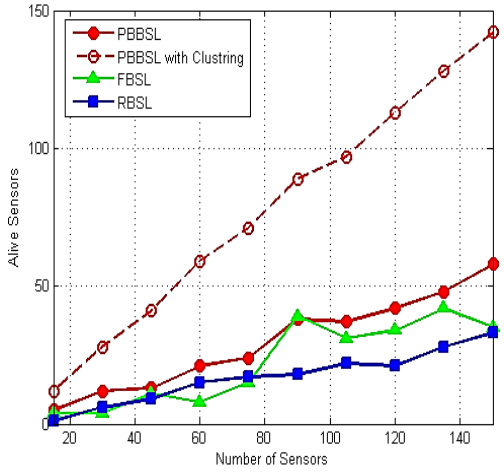


Fig. 4. Number of alive sensors versus different sensors

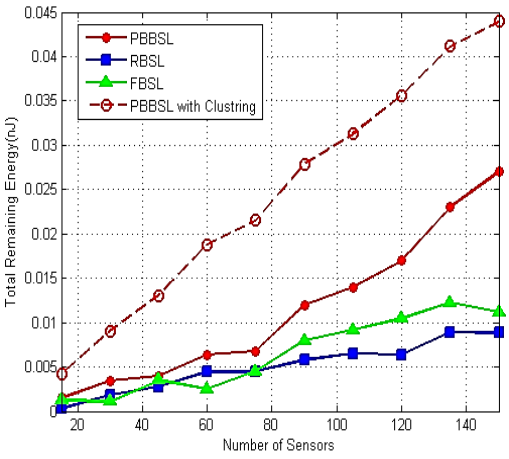


Fig. 5. Total remaining energy versus different sensors

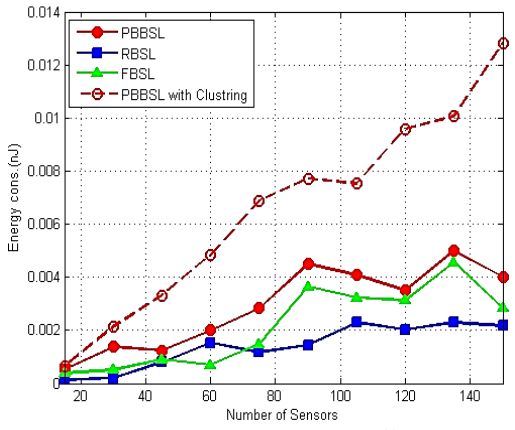


Fig. 6. Total energy consumption versus different sensors

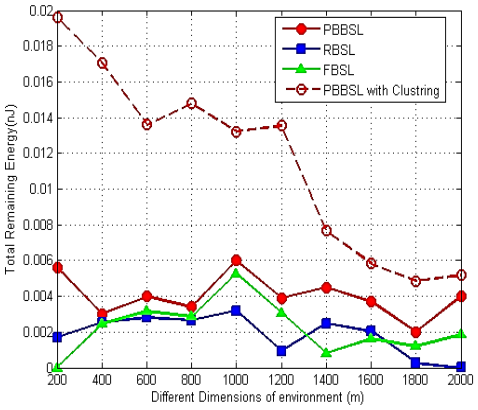


Fig. 7. Total remaining energy versus different environments

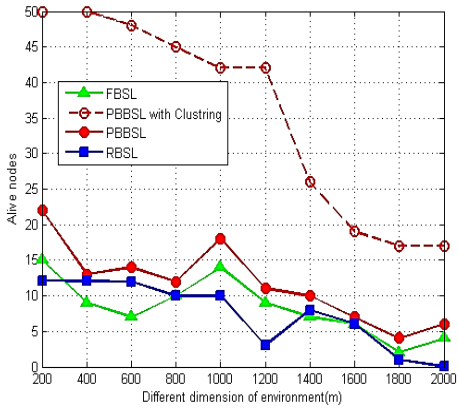


Fig. 8. Number of alive sensors versus different environments

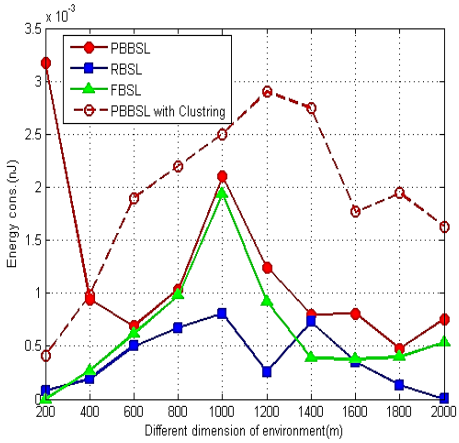


Fig. 9. Total energy consumption versus different environments

## 6. Conclusions

Wireless sensor networks have multiple applications in intelligent environment and structural monitoring. However, in wireless sensor networks, one of the most critical challenges is the power constraint of the sensors. In this paper, we proposed an algorithm based on PSO algorithm to improve the lifetime of the network. For this purpose, at first, the cluster heads are selected among the sensors according to their remaining energy and distances from other nodes. Then, the suitable position of BS is obtained based on the cluster heads position using PSO algorithm. By the proposed algorithm, the energy consumption

of the nodes are more saved and the residual energy is increased. It means that the sensors have more opportunity to be alive and monitor the environment. The simulation results showed the effectiveness of the proposed algorithm in lifetime improvement in different situations. Cluster head selection using the other algorithms and using the mobile sensor networks can be applied as the future work of this paper.

## References


[1] W. B. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy Efficient Communication Protocol for Wireless Microsensor Networks", in Proceedings Sciences, 2000, pp.1-10.

[2] L. Xiang, J. Luo, and A. Vasilakos, " Compressed Data Aggregation for Energy Efficient Wireless Sensor Networks", in 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Adhoc Communications and Networks (SECON), 2011, pp. 46–54.

[3] X. Y. Liu, Y. Zhu , L. Kong, C. Liu, Y. Gu, A. V. Vasilakos, and M.-Y. Wu, "CDC: Compressive Data Collection for Wireless Sensor Networks", IEEE Transactions on Parallel and Distributed Systems, Vol.26, No.8, 2015, pp. 2188–2197.

[4] X. Xu, R. Ansari, A. Khokhar, and A.V. Vasilakos, "Hierarchical Data Aggregation Using Compressive Sensing (HDACS) in WSNs", ACM Transactions on Sensor Networks (TOSN), Vol.11, No.3, 2015, pp.45-45.

[5] W. B. Heinzelman, A.P. Chandrakasan, and H. Balakrishnan, "An Application Specific Protocol Architecture for Wireless Microsensor Networks", IEEE Transactions on Wireless Communications, Vol.1, No.4,2002, pp.660–670.

[6] N. M. A. Latiff, C. C. Tsimenidis, and B. S. Sharif, "Energy-Aware Clustering for Wireless Sensor Networks Using Particle Swarm Optimization", in Proceedings of 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications,2007, pp. 1–5.

[7] B. Singh, and D.K. Lobiyal, "A Novel Energy-Aware Cluster Head Selection Based on Particle Swarm Optimization for Wireless Sensor Networks", Human-Centric Computing and Information Sciences Journal, Vo.l2, No.1, 2012 , pp.2–13.

[8] J. Roselin, P. Latha and S. Benitta, "Maximizing the Wireless Sensor Networks Lifetime through Energy Efficient Connected Coverage", Elsevier Adhoc Networks Journal, Vol.62, 2017, pp.1-10.

[9] S. Arjunan and P. Sujatha, "Lifetime maximization of wireless sensor network using fuzzy based unequal clustering and ACO based routing hybrid protocol", Applied Intelligence springer Journal, Vol.48, No.8, 2018, pp. 2229–2246.

[10] S. Jabbar, M. Ahmad, K.R. Malik, Sh. Khalid, J. Chaudhry and O. Aldabbas, "Designing an energy-aware mechanism for lifetime improvement of wireless sensor networks: a comprehensive study", Mobile Networks and Applications Springer Journal, Vol. 23, No.3, 2018, pp. 432–445.

[11] W. Y. Poe and J. B. Schmitt, "Minimizing the Maximum Delay in Wireless Sensor Networks by intelligent sink placement", Tech. Rep. 362/07, Distributed Computer Systems Lab, University of Kaiserslautern, Kaiserslautern, Germany, 2007.

[12] J. Pan, L. Cai, Y. T. Hou, Y. Shi and S. X. Shen, "Optimal Base Station Locations in Two-Tiered Wireless Sensor Networks", IEEE Transactions on Mobile Computing, Vol. 4, No. 5, 2005, pp. 458-473.

[13] M. 1. Showkat, B. Paul, M. A. Matin, and M. S Alam, "Optimal Sink Location in Wireless Sensor Networks Using Particle Swarm Optimization", in Proc. IEEE Interntional Conference on Antennas, Propagation and Systems (I A009), .Ihor Bahru, Malaysia, 2009, pp. 5445 – 5450.

[14] A. Ebrahimzadeh, M.Najimi, S. M. Hosseni Andargoli, and A. Fallahi, "Sensor Selection and Optimal Energy Detection Threshold for Efficient Cooperative Spectrum Sensing", IEEE Transaction on Vehicular Technology Journal, Vol.64, No.4, 2015, pp. 1565 – 1577.

[15] N. Aslam, and W. Phillips, W. Robertson and Sh. Sivakumar, "A Multi- Criterion Optimization Technique for Energy Efficient Cluster Formation in Wireless Sensor Networks", Information Fusion Journal in Press, Elsevier, Vol. 12. No.3, 2011, pp.202-212.

[16] M. Najimi, A. Ebrahimzadeh, S.M. Hosseni Andargoli, and A. Fallahi, "Lifetime Maximization in Cognitive Sensor Networks Based on the Node Selection", IEEE sensors Journal, Vol. 14, No. 7, 2014, pp.2376-2383.

[17] J. Kennedy, and R. Eberhart, "Particle Swarm Optimization", IEEE International Conference on Neural Networks, 1995, pp. 1942–1948.

[18] M. Azharuddin and P.K. Jana, "Particle Swarm Optimization for Maximizing Lifetime of Wireless Sensor Networks", Computers and Electrical Engineering Journal, Elsevier, Vol.51, 2016, pp.26-42.

[19] R.V. Kulkarni, and G.K. Venayagamoorthy, "Particle Swarm Optimization in Wireless Sensor Networks: A Brief Survey", IEEE Transactions on Systems, Vol.41, No.2, 2011, pp. 262-267.

[20] I.S. Akila, R. Venkatesan, and R. Abinaya, "A PSO Based Energy Efficient Clustering Approach for Wireless Sensor Networks", in IEEE International Conference on Computation of Power, Energy Information and Communication (ICCPEIC), 2016, pp.259-264.

[21] S. Maleki, A. Pandharipande, and G. Leus, "Energy-efficient distributed spectrum sensing for cognitive sensor networks", in Proceedings of 35th Annual Conference IEEE Industrial Electronics, 2009, pp. 2642–2646.

[22] S. Maleki, A. Pandharipande, and G. Leus, "Energy efficient distributed spectrum sensing with convex optimization", in Proceedings of 3rd International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2009, pp. 396–399.

[23] M. Najimi, A. Ebrahimzadeh, S.M. Hosseni Andargoli, and A. Fallahi, "A Novel Sensing Nodes and Decision Node Selection Method for Energy Efficiency of Cooperative Spectrum Sensing in Cognitive Sensor Networks", IEEE Sensors Journal, Vol.13, No.5, 2013, pp.1610-1621.



**Maryam Najimi** received her B.Sc in electronic engineering from Sistan & Baloochestan University; Iran in 2004 and her M.Sc in telecommunication systems engineering from K.N.Toosi University of Thechnology and Ph.D degree in communication engineering from Babol university of Technology. She is currently an assistant professor with department of electrical and computerr engineering, University of Science and Technology of Mazandaran. Her interests include spectrum sensing in cognitive sensor networks.

**Sajjad Nankhoshki** received her B.Sc in electronic engineering from University of Science and Technology of Mazandaran. Her interest includes energy efficiency in wireless sensor networks.

# Using Discrete Hidden Markov Model for Modelling and Forecasting the Tourism Demand in Isfahan

Khatereh Ghasvarian Jahromi*
Department of Electrical Engineering, ACECR Institute of Higher Education (Isfahan Branch), Isfahan, Iran
ghasvarian@jdeihe.ac.ir
Vida Ghasvarian Jahromi
Department of Tourism Management, Faculty of Humanities, University of Science and Arts, Yazd, Iran
v.ghosoorian@stu.sau.ac.ir

Received: 16/06/2018 Revised: 30/10/2018 Accepted: 04/11/2018

**Abstract**

Tourism has been increasingly gaining acceptance as a driving force to enhance the economic growth because it brings the per capita income, employment and foreign currency earnings. Since tourism affects other industries, in many countries, tourism is considered in the economic outlook. The perishable nature of most sections dependent on the tourism has turned the prediction of tourism demand an important issue for future success. The present study, for the first time, uses the Discrete Hidden Markov Model (DHMM) to predict the tourism demand. DHMM is the discrete form of the well-known HMM approach with the capability of parametric modeling the random processes. MATLAB Software is applied to simulate and implement the proposed method. The statistic reports of Iranian and foreign tourists visiting Isfahan gained by Iran Cultural Heritage, Handicrafts, and Tourism Organization (ICHHTO)-Isfahan Tourism used for simulation of the model. To evaluate the proposed method, the prediction results are compared to the results from Artificial Neural Network, Grey model and Persistence method on the same data. Three errors indexes, MAPE (%), RMSE, and MAE, are also applied to have a better comparison between them. The results reveal that compared to three other methods, DHMM performs better in predicting tourism demand for the next year, both for Iranian and foreign tourists.

**Keywords:** Modeling; Tourism Demand Function; Demand Prediction; Discrete Hidden Markov Model; Iran; Isfahan.

## 1. Introduction

For many years and in different parts of the world, tourism has been a driving force for boosting the economic growth through increasing the per capita income, employment rate and foreign currency earnings. Fulfilling these objectives, however, requires appropriate investment in both public and private sectors [1]. Since the tourism affects other industries, in many countries, whether developed or developing, tourism has gained great acceptance in the economic outlook and has increased the development of many other sectors such as agriculture, handicraft, beverages, transportation, etc. [2]. The perishable nature of most related sectors to the tourism has turned the prediction to a very important issue for future success. The long-term and short-term predictions, however, are important for different management objectives; for example, the long-term predictions of tourism demand for next following years help the tourism infrastructure planning in the destination, while short-term demand predictions help the destination flexibility for the next two or three months [3]. Precise estimation of tourism demand helps the tourism managers and industry decision makers in the destination to have a better strategic planning. Hence, in recent years, the prediction of tourism demand has been considered by a number of researchers so that the prediction methods are increasingly being introduced [4]. The superior prediction models are identified based on the features and data used in different studies and help the experts to choose the better prediction methods; finally, it would result in commercial decision makings and effective policies [5].

Regarding the tourist attractions, Iran is among the top ten countries and can gain advantages by the tourism. This is particularly important for the countries that their economies are highly dependent on single-product export (oil) because the sustainable tourism has the unique potential of direct injection of money to the economic cycle. On the other hand, employment on the base of tourism doesn't need the high level of skill and training and it can encompass all level of the society. All the mentioned points express the importance of sustainable tourism growth particularly through economy perspective, attention to the infrastructures and planning in this domain; however, the success key in the planning is the prediction of tourism demand and attempt for increasing this demand.

## 2. Literature Review

Tourism demand forecasting has been an interesting subject for many research studies on the tourism and hosting. Song and Li (2008) examined the proposed methods for tourism demand prediction in recent decades and they found out that prediction techniques usually consist of time series, econometric models, artificial

* Corresponding Author

intelligence approaches, and hybrid methods [6]. The time series models predict the tourist arrivals according to the historical patterns. Most of the research studies have used the time series models to predict and analyze the tourism demand [3],[7]-[9]. The most popular model among them is the Autoregressive Moving Average Model [6]. The econometric models examine the cause-effect relationship between incoming tourists and the effective factors [10] and this model is particularly useful when there is a correlational relationship between the factors. The artificial intelligent methods use the neural networks and vector machines for the nonlinear data modeling [11],[12]. Some research studies have proposed the hybrid methods by combining data mining and econometric models [4],[13]. The researchers have also used the meta-analysis and singular spectrum analysis in modeling and prediction of tourists visiting a place [5],[14]. Regarding the prediction exactness, different models have some advantages and disadvantages. No single model can be steadily superior to other models in all situations [6]. The artificial intelligence techniques can only model limited observations. For example, Wang (2004) used fuzzy time series models to estimate the tourism demand using 12 data points [15]. The econometric models need a large number of observations to have higher precision in prediction; in comparison, the artificial intelligence models have no theoretical bases for modeling the tourism demand and the researchers are not able to show the partial effects of each explanatory variable on another explanatory variable [6]. In contrast, the econometric models have an appropriate theoretical basis and can confirm the relations between the explanatory variables through the economic perspective. Other quantitative methods such as gravity models, artificial neural networks (ANNs) and single-variable time series models have also an important role in the tourism demand prediction.

In recent years, researchers have paid attention to artificial intelligence (AI) and hybrid methods in addition to studying on improving accuracy. Li, Song, and Shen (2011) evaluated six hybrid methods for prediction of British outbound tourism demand in seven destination countries [16]. Their numerical results show that generally speaking, the hybrid methods yield better results than single techniques. Chen (2011) combined the linear and non-linear statistical models to predict the foreign tourists in Taiwan [17]. The empirical results showed that the hybrid Support Vector Regression (SVR) models can detect the directional change. Peng et al. (2014) reviewed 65 published studies from 1980 to 2011; they examined the accuracy of different prediction models, data features, and the conditions of each study by meta-regression analysis [5]. They showed that the origin and destination of the tourists, time period, modeling method, data alternation, number of variables, variables measurement, and the data volume considerably affect the prediction accuracy. Cheng and Liu (2014) compared gray forecasting model and cubic polynomial for Guilin tourism demand data [18]. They proposed a hybrid prediction model to improve prediction accuracy. Wang, Zhang, and Guo (2015) used the synthetic index method to calculate the tourism market growth index in order to achieve the annual tourism forecast [19]. The sample data were trained using the machine learning algorithm. Ultimately, they obtained a model based on Extreme Learning Machine (ELM) for forecasting tourism demand in Liaoning province and compared the results with the SVR algorithm. Liang (2016) combined the autocorrelation function (ACF), neural networks, and genetic algorithms to forecast tourism demand [20]. They compared the hybrid model with neural networks and the Seasonal Autoregressive Integrated Moving Average (SARIMA) models in forecasting Taiwan's tourism demand from 2001 to 2009. Sadati, Bateni, and Bateni (2016) examined the effectiveness of ANNs as an alternative approach to the use of SVR in the tourism research [21]. They evaluated the method by prediction the tourism demand in Iran. Sun et al. (2016) used a new prediction model called Cuckoo-Markov Chain-Segment Grey model (1,1) to evaluate the prediction accuracy undergoing tourism market fluctuations [13]. They showed that this method is considerably more efficient and precise than the usual Markov Chain-Grey models (1,1). Rossello and Sanso (2017) found out that the entropy and relative abundance are quite appropriate as the seasonal indexes and can be applied as a new information tool for the seasonal analysis of tourism [22]. Kazak (2018) used the statistical and econometric modeling of tourist expenses to evaluate and forecast tourism development [23]. He showed that the dynamics of demand for tourism can be affected by a wide range of factors, such as political factors and economic relations.

Although many studies in forecasting tourism demand are based on the combination of the previous methods, some researchers use the new methods in this area. Li and Cao (2018) used Short-Term Memory Neural Networks (LSTM) to predict the flow of tourism [24]. They showed that this method performs better than the ARIMA model and the Back Propagation Neural Network (BPNN) on the data obtained from the Xi'an Museum. Yao et al. (2018) presented a new model of the neural networks [25]. In their proposed model, tourist arrival data were decomposed by two low-pass filters into a long-term trend and short-term seasonal components and then modeled by a pair of autoregressive neural networks as a parallel structure. This method was evaluated by the data of tourist arrival to United States from twelve markets. Sun et al. (2019) proposed a prediction framework that used machine learning and internet search indexes to predict the arrival of tourists to popular destinations in China [26]. They compared the proposed model performance with Google and Baidu's search results. The study confirmed the Granger causality and co-integration relationship between the Internet search index and the arrival of tourists to Beijing.

In the present study, in order to forecasting the tourism demand, the Discrete Hidden Markov Model (DHMM) is used for the first time.

## 3. Theoretical Framework

The present study used the discrete type of Hidden Markov model (HMM) to predict the tourism demand. HMM is a statistical modeling tool for time series that has been applied successfully in the speech recognition, error detection, and computational processing. HMM performs statistical analysis and parametric modeling on unstable signals. That is why it can simply be used for probability-based reasoning. This property of HMM has been used to predict tourism demand in the present study.

### 3.1 Introduction on the Discrete Hidden Markov Model (DHMM)

Based on the observed values, two types of HMMs have been introduced: Continuous Hidden Markov Model (CHMM) in which the observed values are continuous; and Discrete Hidden Markov Model (DHMM) in which the sequence of the observed values are in a defined range of codes or symbols. HMM has two parts: Markov chain and random process. The Markov chain with a sequence of states as the output is described by a vector $\pi$ and matrix A, and the random process with observed values as the output is described by matrix B [27]. Fig. 1 shows the HMM structure in which T is the length of time sequence.


Fig. 1. The HMM Structure [21]

A DHMM is described with the following parameters [28]:

- $N$: Number of Markov chain states; if $\theta_1, \theta_2, \ldots, \theta_N$ are possible states of the Markov chain and $q_t$ is the state at time $t$, then $q_t \epsilon (\theta_1, \theta_2, \ldots, \theta_N)$.
- $M$: Number of observed values in each state; if $v_1, v_2, \ldots, v_M$ are the observed values and $o_t$ is the value at time $t$, then $o_t \epsilon (v_1, v_2, \ldots, v_M)$.
- $\Pi$: Initial probability distribution vector, in which:

$$\pi_i = P(q_i = \theta_i), \qquad 1 \le i \le N \qquad (1)$$

- $A$: State transition probability matrix, $A = (a_{ij})_{N \times N}$, in which:

$$a_{ij} = P(q_{t+1} = \theta_j / q_t = \theta_i), \quad 1 \le i \le N \qquad (2)$$

- $B$: Observation probability matrix, $B = (b_{jk})_{N \times M}$, in which:

$$b_{jk} = P(o_t = v_k / q_t = \theta_j), \quad 1 \le j \le N, \quad 1 \le k \le M \qquad (3)$$

In short, a DHMM can be expressed in form $\lambda = (N, M, \pi, A, B)$.

### 3.2 The HMM-based Prediction Procedure

The overall HMM-based prediction procedure consists of two steps:

I. Training Process:

In the training phase, the parameters of $\lambda = (\pi, A, B)$ are trained according to the model. The parameters keep updating until the best adaptation with the model is obtained. This process needs the Baum-Welch algorithm that uses the well-known Expectation-Maximization (EM) algorithm to improve the likelihood $P(O/\lambda)$ or log-likelihood $ln\, P(O/\lambda)$ [23]. To apply this algorithm, an initial guess of matrices $A$ and $B$ is required.

II. Decoding Process:

In this stage, the probability (or probability logarithm) of the observation of a sequence is calculated by trained HMM. To do so, the forward-backward algorithm is used [24]. The sequence with the highest probability of observation is the base of the prediction.

Next section introduces more explanations on the way DHMM is used to predict the tourism demand.

## 4. Using DHMM for Predicting the Tourism Demand in the Next Year in Isfahan

Fig. 2 and Fig. 3 graphically illustrate the monthly and yearly statistics of Iranian and foreign tourists checked-in to the hotels in Isfahan since 2002 to 2016. These figures were gained by Iran Cultural Heritage, Handicrafts, and Tourism Organization (ICHHTO)-Isfahan Tourism. As Fig. 2 and Fig. 3 show, the number of Iranian tourists in April, September, and August is more and the number of foreign tourists is more in May, October, and August. Generally speaking, it can be inferred that some times of the year (month) are considered as the popular time of visiting. More or less, all the tourist destinations and business centers are facing the seasonal nature of the tourism and this can be a reason explaining why the population of tourists and visitors in a destination is so variable in different seasons. That is why the tourist destinations sometimes are too crowded and populated to meet the needs of the tourists and in some other times, some businesses are facing considerable recession.

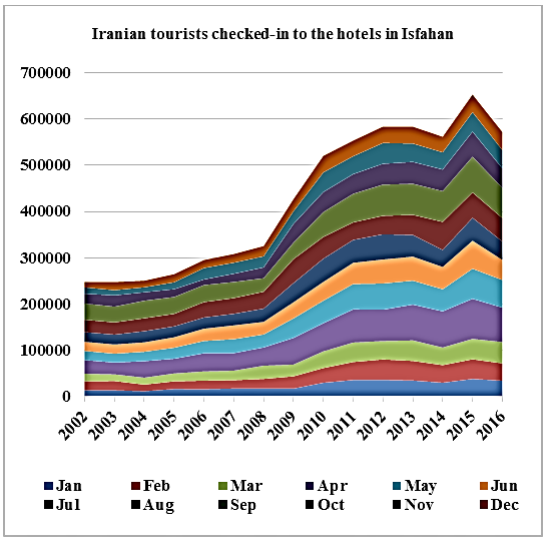


Fig. 2. The number of Iranian tourists checked-in to the hotels in Isfahan since 2002 to 2016

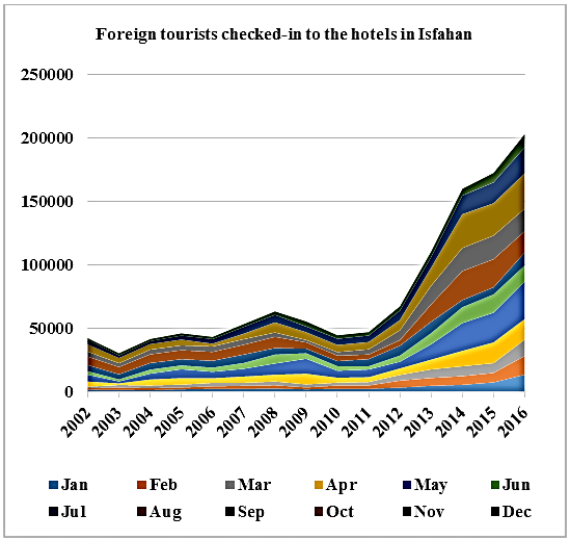


Fig. 3. The number of foreign tourists checked-in to the hotels in Isfahan since 2002 to 2016

Since prediction in tourism is one of the measuring methods for tourism demand, preparing the infrastructures and amenities both in long term and short term, and taking into the seasonal nature of tourism, the monthly periods must be compared during consecutive years rather than a general comparison of the yearly demand. To do so, in the proposed method, the tourism statistic in a month (for example April) in the previous years was used to predict the tourism demand of the same month.

To predict the monthly tourism demand in 2016, the monthly statistics of 2002 to 2015 was used to train the DHMM. The year 2016 was chosen because it has a relatively unpredictable statistic particularly regarding Iranian tourists and its real statistic is also available and comparable. To simulate the proposed method, the MATLAB software was used.

Considering the objective- the monthly prediction of 2016-, first the monthly statistic growth of each year (2003 to 2015) relative to the previous year is measured; then the gained values are normalized, coded and used for training the DHMM (for each value a code is taken). Since the tourism demand for each month is predicted according to the same month in the previous year, 12 DHMM models are required to predict 12 months in each mode of Iranian and foreign tourists (24 models in total).

The required parameters for DHMM training are determined as the following:

- The number of Markov chain states (hidden states) is N = 3.
- The number of possible observed values in each state (M) depends on the number of codes. This value is M=702 and M=1865 for Iranian and foreign tourists, respectively.
- The initial guess of the matrix A for use in the Baum-Welch algorithm is chosen randomly for 24 models.
- For the initial guess of matrix B in the Baum-Welch algorithm, a matrix with identical elements is considered, which, given the size of the matrix (N*M), would be different for Iranian and foreign tourists.

For DHMM training the sequences of length 10, indicating the value changes in 10 consecutive years, are considered.

After training the DHMM by the mentioned data, the final matrices A and B are obtained for each model.

In the decoding step, the probability of occurrence for each observable value is measurable. In this stage, the occurrence probability of different 10-element sequences are examined and the corresponding values are predicted according to the codes with the highest probability. For example, to predict the demand for Iranian tourists in April 2016, firstly the corresponding code must be predicted. To do so, the codes of April in the last 9 years are used. Considering the training of each 24 models of the DHMM by 10-element sequences (related to 10 consecutive years), the highest probability among the existing codes would be obtained for the 10th year. This code helps to calculate the tourism demand for the year 2016.

Table 1. Predicted values of the Iranian tourists in 2016 by DHMM in comparison with the actual values, Persistence, Grey, and ANN methods

| | *Real* | *Persistence* | *Grey(1,1)* | *ANN* | *DHMM* |
|---|---|---|---|---|---|
| *Jan* | 34657 | 38506 | 44558 | 53056 | 38256 |
| *Feb* | 37519 | 42451 | 51862 | 43951 | 43151 |
| *Mar* | 45911 | 43839 | 53210 | 59389 | 45689 |
| *Apr* | 74439 | 86777 | 99555 | 80827 | 82927 |
| *May* | 59603 | 64887 | 71465 | 61737 | 64687 |
| *Jun* | 43414 | 60341 | 66601 | 68841 | 57691 |
| *Jul* | 39687 | 49566 | 57334 | 44066 | 51716 |
| *Aug* | 50490 | 54836 | 59376 | 55386 | 50886 |
| *Sep* | 65178 | 77074 | 82758 | 78224 | 71324 |
| *Oct* | 42130 | 54242 | 63100 | 47542 | 53342 |
| *Nov* | 39558 | 42052 | 52978 | 41802 | 41652 |
| *Dec* | 39117 | 38096 | 44545 | 31446 | 38696 |
| *Total* | 571703 | 652667 | 747341 | 666267 | 640017 |
| *MAPE(%)* | | 15.24 | 30.72 | 19.22 | 12.17 |
| *RMSE* | | 8766 | 15851 | 11402 | 7413 |
| *MAE* | | 7263 | 14637 | 9159 | 5800 |

Table 2. Predicted values of the foreign tourists in 2016 by DHMM in comparison with the actual values, Persistence, Grey, and ANN methods

| | *Real* | *Persistence* | *Grey(1,1)* | *ANN* | *DHMM* |
|---|---|---|---|---|---|
| *Jan* | 13781 | 7749 | 7085 | 9659 | 8429 |
| *Feb* | 14684 | 7421 | 8088 | 6491 | 10171 |
| *Mar* | 12957 | 7668 | 8169 | 9458 | 8768 |
| *Apr* | 16118 | 16294 | 13540 | 15854 | 16244 |
| *May* | 29696 | 23384 | 21154 | 21854 | 26094 |
| *Jun* | 12489 | 14337 | 12622 | 15397 | 13777 |
| *Jul* | 10144 | 5666 | 6952 | 11416 | 7776 |
| *Aug* | 16731 | 22025 | 17402 | 14255 | 21635 |
| *Sep* | 17763 | 18564 | 7084 | 19614 | 17554 |
| *Oct* | 27807 | 25490 | 9944 | 39480 | 27840 |
| *Nov* | 20419 | 16376 | 15553 | 12326 | 17626 |
| *Dec* | 10381 | 7535 | 6974 | 8645 | 8335 |
| *Total* | 202970 | 172509 | 134568 | 184449 | 184249 |
| *MAPE(%)* | | 23.01 | 34.49 | 26.57 | 15.48 |
| *RMSE* | | 4462 | 7472 | 5645 | 3200 |
| *MAE* | | 3892 | 5834 | 4494 | 2619 |

Using this method, the demand for Iranian and foreign tourism in Isfahan for 12 months in 2016 was predicted. Table 1 and Table 2 indicate the predicted values in this year whereas 12 DHMM models were used to obtain each of them. For comparison, the prediction results were obtained from the Artificial Neural Network (ANN), Grey Model (1,1) and Persistence method in addition to DHMM. ANN is a well-known artificial intelligence

method, which is used in many prediction models, Grey model is a popular model for researchers due to its ability to track the fluctuation of observations [29], and finally Persistence Method is also a benchmark model in which the previous latest observed value is considered as the prediction of the next value [30].

To determine the exactness of the proposed method, the Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) were used, which are obtained by the following relationships:

$$MAPE = \frac{100}{N}\sum_{t=1}^{N}\frac{\left|x_p(t)-x_r(t)\right|}{\overline{x_r}} \quad (4)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(x_p(t)-x_r(t)\right)^2} \quad (5)$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|x_p(t)-x_r(t)\right| \quad (6)$$

$$\bar{x} = \frac{1}{N}\sum_{t=1}^{N}x_r(t) \quad (7)$$

where $x_r(t)$ and $x_p(t)$ are the actual and predicted values, respectively.

As it is shown in Table 1 the MAPE index gained by DHMM is 12.17 percent, while this index for ANN, Grey and Persistence methods is 19.22, 30.72 and 15.24 percent, respectively. Also, according to Table 2, this index for DHMM, ANN, Grey, and Persistence methods are 15.48, 26.57, 34.49, and 23.01 percent, respectively indicating less error in the proposed method. On the other hand, comparing the RMSE and MAE indexes in Table 1 and Table 2 suggests that DHMM is superior to the three other methods. Given the fact that in 2016 Iranian tourism had a negative growth rate in comparison to 2015 (Fig. 2), the lower error in DHMM prediction suggests the higher ability of this method. It should be mentioned that the more trained data in DHMM, the higher ability in prediction would be achieved.

Fig. 4 and Fig. 5 show the curves of predicted Iranian and foreign tourism demand in Isfahan for 12 months in 2016 by using DHMM as well as the actual values and predicted

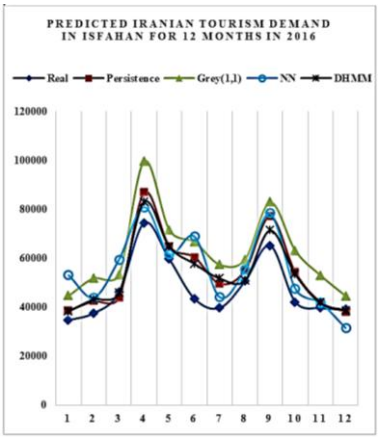


Fig. 4. Comparison between prediction by DHMM and three other methods in Iranian tourism demand

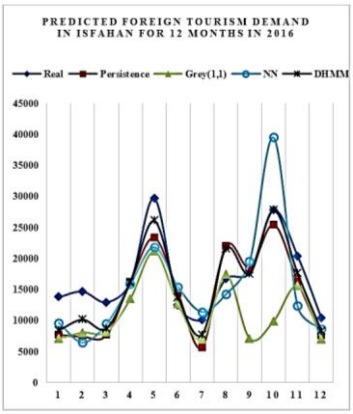


Fig. 5. Comparison between prediction by DHMM and three other methods in foreign tourism demand

values by three other methods. Furthermore, Fig. 6 and Fig. 7 illustrate the absolute errors of four prediction methods. The total absolute errors for Iranian tourism prediction by Persistence, Grey, ANN, and DHMM are 87150, 175638, 109906, and 69600, respectively; moreover, these errors for foreign tourism prediction are 46699, 70011, 53929, and 31423, respectively. These values show that for both Iranian and foreign tourists the total forecast absolute errors of the suggested method are less than three other methods.

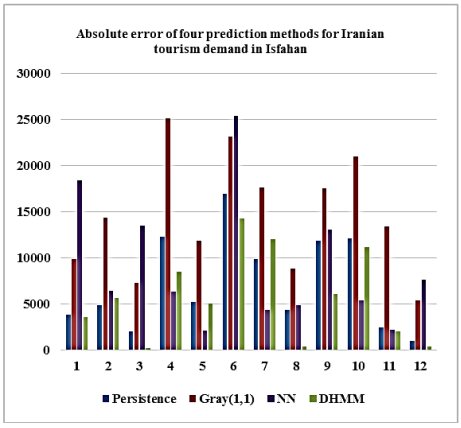


Fig. 6. Comparison between absolute errors of four prediction methods in Iranian tourism demand

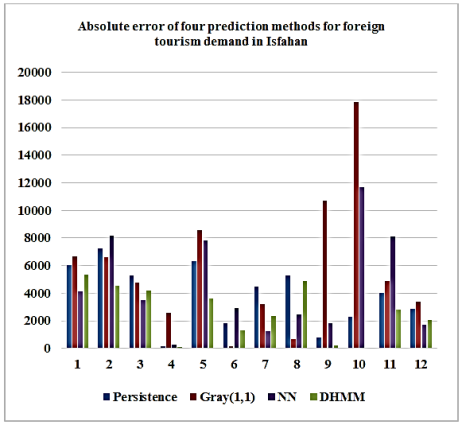


Fig. 7. Comparison between absolute errors of four prediction methods in foreign tourism demand

## 5. Conclusion

In this research, the Discrete Hidden Markov Model (DHMM) was used to predict the tourism demand in Isfahan. This is the first time that one of the hidden Markov models is being used to predict the demand for tourism. The DHMM was simulated in MATLAB software and it was implemented on the statistic of Iranian and foreign tourists visiting Isfahan. To determine the efficiency of the proposed method, its results were compared with the results of the NN model, grey model, and persistence method. The error rate of these methods was obtained by using three error indexes. The results also showed that DHMM performs more satisfactory than the other mentioned methods; besides, less error in the proposed method suggests a more realistic prediction in tourism demand. Since the state and non-profit organizations are involved in the tourism industry, the tourism demand is simply influenced by social, political, economy, cultural events. Thus, a method with less error in sudden changes is more reliable regarding the prediction. The managers and decision-makers in the tourism industry can enjoy this method to plan the short-term facilities and improve infrastructures in the long-term. The researchers also can use the DHMM as a basic method in producing the hybrid methods for predicting the tourism demand.

## References


[1] F. L. Chu, "Using a logistic growth regression model to forecast the demand for tourism in Las Vegas," Tour. Manag. Perspect., vol. 12, pp. 62–67, 2014.

[2] "Strategic Plan for the Development of Tourism in Mozambique (2004-2013)," 2013.

[3] U. Gunter and I. Önder, "Forecasting international city tourism demand for Paris: Accuracy of uni- and multivariate models employing monthly data," Tour. Manag., vol. 46, pp. 123–135, 2015.

[4] P. F. Pai, K. C. Hung, and K. P. Lin, "Tourism demand forecasting using novel hybrid system," Expert Syst. Appl., vol. 41, no. 8, pp. 3691–3702, 2014.

[5] B. Peng, H. Song, and G. I. Crouch, "A meta-analysis of international tourism demand forecasting and implications for practice," Tour. Manag., vol. 45, pp. 181–193, 2014.

[6] H. Song and G. Li, "Tourism demand modelling and forecasting-A review of recent research," Tour. Manag., vol. 29, no. 2, pp. 203–220, 2008.

[7] M. Akin, "A novel approach to model selection in tourism demand modeling," Tour. Manag., vol. 48, pp. 64–72, 2015.

[8] F. L. Chu, "Analyzing and forecasting tourism demand with ARAR algorithm," Tour. Manag., vol. 29, no. 6, pp. 1185–1196, 2008.

[9] A. Guizzardi and A. Stacchini, "Real-time forecasting regional tourism with business sentiment surveys," Tour. Manag., vol. 47, pp. 213–223, 2015.

[10] H. Song, G. Li, S. Witt, and B. Fei, "Tourism demand modelling and forecasting: how should demand be measured?," Tour. Econ., vol. 16, no. 1, pp. 63–81, 2010.

[11] A. Palmer, J. José Montaño, and A. Sesé, "Designing an artificial neural network for forecasting tourism time series," Tour. Manag., vol. 27, no. 5, pp. 781–790, 2006.

[12] E. Hadavandi, A. Ghanbari, K. Shahanaghi, and S. Abbasian-Naghneh, "Tourist arrival forecasting by evolutionary fuzzy systems," Tour. Manag., vol. 32, no. 5, pp. 1196–1203, 2011.

[13] X. Sun, W. Sun, J. Wang, Y. Zhang, and Y. Gao, "Using a Grey-Markov model optimized by Cuckoo search algorithm to forecast the annual foreign tourist arrivals to China," Tour. Manag., vol. 52, pp. 369–379, 2016.

[14] H. Hassani, A. Webster, E. S. Silva, and S. Heravi, "Forecasting U.S. Tourist arrivals using optimal Singular Spectrum Analysis," Tour. Manag., vol. 46, pp. 322–335, 2015.

[15] C. H. Wang, "Predicting tourism demand using fuzzy time series and hybrid grey theory," Tour. Manag., vol. 25, no. 3, pp. 367–374, 2004.

[16] S. Shen, G. Li, and H. Song, "Combination forecasts of International tourism demand," Ann. Tour. Res., vol. 38, no. 1, pp. 72–89, 2011.

[17] K. Y. Chen, "Combining linear and nonlinear model in forecasting tourism demand," Expert Syst. Appl., vol. 38, no. 8, pp. 10368–10376, 2011.

[18] D. Cheng and L. Bin Liu, "Forecasting of tourism demand for Guilin based on combined model," in 7th International Joint Conference on Computational Sciences and Optimization, CSO 2014, 2014, no. 1, pp. 100–103.

[19] X. Wang, H. Zhang, and X. Guo, "Demand Forecasting Models of Tourism Based on ELM," in 7th International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2015, 2015, pp. 326–330.

[20] Y. H. Liang, "Using the combined model for forecasting the tourism demand," in International Conference on Machine Learning and Cybernetics, 2016, vol. 2, pp. 612–615.

[21] S. Sadatiseyedmahalleh, N. H. Bateni, and N. H. Bateni, "Prediction of Tourism Demand in Iran by Using Artificial Neural Network (ANN) and Supporting Vector Machine (SVR)," Int. J. Multicult. Multireligious Underst., vol. 3, no. 1, pp. 37–44, 2016.

[22] J. Rosselló and A. Sansó, "Yearly, monthly and weekly seasonality of tourism demand: A decomposition analysis," Tour. Manag., vol. 60, pp. 379–389, 2017.

[23] A. N. Kazak, "Analysis of Dynamics of Demand, Revenue and Ergonomic Aspects of Tourism," in Third International Conference on Human Factors in Complex Technical Systems and Environments (ERGO)s and Environments (ERGO), 2018, pp. 13–15.

[24] Y. Li and H. Cao, "Prediction for Tourism Flow based on LSTM Neural Network," Procedia Comput. Sci., vol. 129, pp. 277–283, 2018.

[25] Y. Yao et al., "A paired neural network model for tourist arrival forecasting," Expert Syst. Appl., vol. 114, pp. 588–614, 2018.

[26] S. Sun, Y. Wei, K. L. Tsui, and S. Wang, "Forecasting tourist arrivals with machine learning and internet search index," Tour. Manag., vol. 70, pp. 1–10, 2019.

[27] J. Kang, C. Feng, Q. Shao, and H. Hu, "Prediction of Chatter in Machining Process Based on Hybrid SOM-DHMM Architecture," in ICIC 2007, LNAI 4682, 2007, pp. 1004–1013.

[28] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in Proceedings of the IEEE, 1989, vol. 77, no. 2.

[29] A. Asrari, D. Seyed Javan, M. Hossein Javidi, and M. Monfared, "Application of Gray-Fuzzy-Markov Chain Method for Day-Ahead Electric Load Forecasting," Przegląd Elektrotechniczny Electr. Rev., no. 3, pp. 228–237, 2012.

[30] N. Amjady, F. Keynia, and H. Zareipour, "Wind Power Prediction by a New Forecast Engine Composed of Modi fi ed Hybrid Neural Network and Enhanced Particle Swarm Optimization," IEEE Trans. Sustain. Energy, vol. 2, no. 3, pp. 265–276, 2011.

**Khatereh Ghasvarian Jahromi** received the B.Sc. degree in Control Engineering from Isfahan University of Technology, Isfahan, Iran, in 1998 and the M.Sc. degree in Telecommunication Engineering from Shiraz University, Shiraz, Iran, in 2001. In 2002, she joined the Department of Electrical Engineering at the ACECR Institute of Higher Education (Isfahan Branch) as a faculty member. Currently, she is a Ph.D. student of electrical engineering at Shahid Beheshti University, Tehran, Iran. Her area research interests include Markov chain, Hidden Markov models, genetic algorithms, forecasting models, and statistical analysis methods.

**Vida Ghasvarian Jahromi** received the B.Sc. degree in Industrial Engineering from the Payame Noor University of Isfahan, Isfahan, Iran, in 2012 and the M.Sc. degree in Tourism Management from Science & Art University of Yazd, Yazd, Iran, in 2018. Her area research interests include tourism management, social responsibilities, fuzzy concepts, and statistical analysis methods.