

<https://doi.org/10.7236/IIBC.2017.17.2.75>

IIBC 2017-2-12

블루투스 접촉 데이터를 이용한 사회관계구조 검출 알고리즘

Detection Algorithm of Social Community Structure based on Bluetooth Contact Data

웬공빈*, 윤석훈**

Nguyen Cong Binh*, Seokhoon Yoon**

요약 본 논문에서는 사회관계구조에 초점을 맞춘 사회관계망 분석을 고려한다. 사회관계망은 많은 사회집단으로 구성되어 있으며, 사회관계 구조 특성으로 인하여 같은 사회집단 내의 노드들은 서로 강한 유대관계를 가지고 있으며 다른 사회집단에 속한 노드와는 상대적으로 약한 유대를 가지게 된다. 사회관계망에서의 사회관계구조 검출은 사람들의 행동 및 상호작용의 분석과 예측을 가능하게 한다. 본 논문에서는 사회관계구조와 사회집단을 검출하기 위하여 사람들이 소지하는 스마트기기의 실제 블루투스 접촉 데이터를 이용한다. 네트워크 노드 간 유대를 추정하기 위한 다양한 유사도 측정 방식과 클러스터링을 기반으로 하는 사회관계구조 검출 방안을 제시한다. 제안하는 방안을 검증하기 위하여 고유관계 특성을 이용하는 성능측정방안을 이용한다.

Abstract In this paper, we consider social network analysis that focuses on community detection. Social networks embed community structure characteristics, i.e., a society can be partitioned into many social groups of individuals, with dense intra-group connections and much sparser inter-group connections. Exploring the community structure allows predicting as well as understanding individual's behaviors and interactions between people. In this paper, based on the interaction information extracted from a real-life Bluetooth contacts, we aim to reveal the social groups in a society of mobile carriers. Focusing on estimating the closeness of relationships between network entities through different similarity measurement methods, we introduce the clustering scheme to determine the underlying social structure. To evaluate our community detection method, we present the evaluation mechanism based on the basic properties of friendship.

Key Words : Clustering, Community structure, Friendship, Similarity Measurement, Social network analysis.

1. Introduction

Social network analysis is the technique that provides the studies about relationships and interactions among network nodes [1-3]. Analyzing human society

has gained significant research interest, since exploring the organization of human networks may help to uncover the patterns of human behaviors and social correlations [4]. Community detection is one promising aspect in human networks analysis. Research has

*준회원, 울산대학교 전기전자컴퓨터공학과

**정회원, 울산대학교 전기전자컴퓨터공학과 (교신저자)

접수일자 2017년 1월 9일, 수정완료 2017년 2월 17일

게재확정일자 2017년 4월 7일

Received: 9 January, 2017 / Revised: 17 February, 2017 /

Accepted: 7 April, 2017

**Corresponding Author: seokhoonyoon@ulsan.ac.kr

Dept. of Electrical and Computer Engineering, University of Ulsan, Korea

shown that a social network could be modeled as a set of communities^[5]. In other words, human society can be partitioned into multiple social groups, in which the connections between members within the same group are much stronger as well as denser than the links with people in different groups. People in the same social group have strong social ties and interdependence; thus they tend to have correlative characteristics and interact more frequently with each other at internal community level. Discovering which communities they belong to may allow understanding of individual's behavior and interactions between network entities, and hence can facilitate cooperation among them, for data routing and dissemination.

In the recent years, mobile phones incorporating many sensing functions, have become indispensable in most people's lives. As mobile phones are often kept in proximity by users, they become essential tools to capture human activities, thus increasing the amount of data source recording the personal behavior and interpersonal interactions is available for investigation.

There has been a lot of network analysis research focused on community detection^[5-6]. One of the well-known studies which employed real-life mobility trace in human social analysis is eigenbehaviors^[7]. Eagle and Pentland compute the social behavioral distance, such as the number of Bluetooth devices seen, between an individual and subjects in predefined communities, and then by comparing those distances, they can derive which community that one belongs to. The authors in [8] introduced the method of estimating the character of friendship between network nodes, with the information extracted from their contact histories. They proposed social pressure metric that reflects the motivation for interactions among nodes. The estimation relies on three properties of friendship including frequency, longevity, and regularity. In [4], the authors analyzed the longitudinal behavioral data from self-report or logged by mobile devices to infer social ties. In [9], the social structure is determined through analysis of communication logs.

In this paper, we aim to detect the communities within a social network. Based on data from a real-life mobility traces, we will determine the underlying structure among a society of mobile carriers. We focus on constructing a social graph that can well represent the social ties between network entities, in other words, determine the social similarity between subjects, by examining their contact histories data, such as Bluetooth contact traces. To be more specific, in compare with the eigenbehaviors based approach^[7], we introduce the method that estimates the pairwise closeness between network nodes, based on counting the number of contacts and estimating the social relationship between subjects. After similarity measurement step, we propose the mechanism of clustering those mobile users into social groups by applying spectral algorithm^[13]. Moreover, to evaluate proposed scheme, we present the friendship-based evaluation method employing common properties of the friendship, such as frequency, longevity, and regularity.

The rest of this paper is organized as follows. Section II describes the real-life mobility dataset that we use in this paper, as well as how the useful data is extracted from them. Section III demonstrates the clustering analysis methodology. In section IV, the evaluation method is explained, and the results and discussion are presented. Finally, we conclude our work in section V.

II. The dataset

To analyze the human social networks, we examine the MIT Reality Mining Dataset^[10]. This dataset was collected in 2004, over the course of nine months. It contains mobility traces, including call logs, Bluetooth devices in proximity, cell tower logs, application usage, which were collected from 94 mobile users in MIT. Most of the participants are gathered from MIT Media lab students and adjacent business school students. The organization of subjects suggests that there are many social connections among participants, thus exist

the underlying community structure.

In this experiment, each subject carries Bluetooth-enabled mobile phone that periodically scans the environment every 5 minutes, providing the list of Bluetooth devices in its proximity. Since participants frequently carry their phones on them, this data could be used to determine the interactions between those mobile users. In the scope of this paper, each Bluetooth contact among mobile devices will be implied as a contact event between their owners. Moreover, it should be emphasized that the number of meetings between subjects may not be identical with the number of contacts. Because one meeting can last longer than one Bluetooth scanning period, so a meeting may be accounted for many contact events. Therefore, we define one meeting as a sequence of consecutive times in contact between two subjects. We use the dataset of 43 subjects that have sufficient data over the experiment period.

In Reality Mining experiment, different subjects have different participated period, and they may have turned off their phones many times during experiment, leading to the lack of analyzing data. The overlapping period, during which chosen subjects turned on their phones almost all the time, is used to obtain the data. In this paper, a chosen overlapping period is from September 23rd to December 7th, 2004. Fig. 1 and Fig. 2 illustrate the data available from set of filtered subjects in overlapping period. It seems that human interactions are mostly concentrated in certain period of daytime, such as from 9h to 18h (Fig. 1). Hence, we focus on analyzing data extracted from that daytime period. Besides, in general, there is a significant difference between human social patterns during weekend and weekday. The daily behaviors tend to be more consistent during weekday than weekend. As can be seen in Fig. 2, there are fewer human interactions during weekend than weekday. Therefore, we only interest in weekday and exclude weekend data from our analysis.

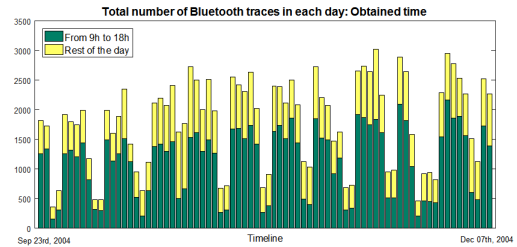


그림 1. 다른 시간대별 블루투스 접촉 수
Fig. 1. Number of Bluetooth traces for each day, obtained in different time periods

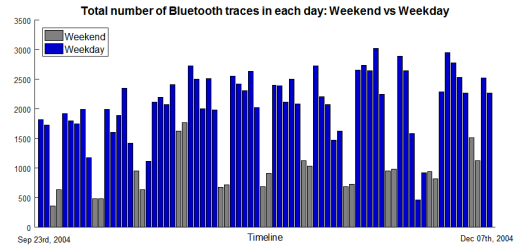


그림 2. 요일별 블루투스 접촉 수
Fig. 2. Number of Bluetooth traces obtained in different days of the week.

III. Social Community Structure Detection Algorithm

In this section, we present the social community structure detection algorithm, which will partition users into different social groups i.e., social groups are detected. As a result of the algorithm, users would have close relations with other group mates in the same social group. The algorithm consists of two main steps, which are similarity measurement and clustering.

Firstly, the similarity between mobile nodes is measured based on their contact histories. The purpose of similarity measurement is to examine the social closeness between subjects. In this step, the level of closeness is estimated using contact history of subjects. The output of this step is the input for the second step of the algorithm, which is clustering to partition the human carried nodes into social groups, whereby nodes in the same group have close relations.

Now, we describe those two steps of the algorithm in more detail.

1. Similarity Measurement

The first step of the scheme is to quantify affinities between subjects. The performance of clustering depends heavily on the similarity metric. In this section, we consider two different similarity measurement methods. In the first method, we measure similarity using behavioral data based on eigenbehavioral similarity, which is extended from eigenbehaviors^[7]. In the second method, we propose a social relationship based similarity. That means, when measuring the relation between subject A and subject B, we take into account the social relationship of subject A with the persons who is socially closest to B.

1.1 Eigenbehavioral similarity

In [7], Eagle and Pentland proposed behavioral similarity measurement between an individual and each subject in a predefined community. For a community (e.g. students in business major), they create behavioral data matrix of m_i by 24, in which m_i is the number of subjects in community i . Each row vector reflects the behavior of individual in community, in this context, it means the number of subjects that one was in contact over the experiment. Those 24 columns correspond to 24 hourly intervals of the day.

In this paper, we extend that approach by considering one large society consisting of all subjects and obtain similarities between them based on PCA^[11] approach. We determine the social distance from an individual to another with no prior knowledge about their social background. Instead of forming behavioral matrix for each community, we create one matrix of M by H to represent all subjects, where $M = \sum m_i$, in our context, $M=43$ (number of subjects) and $H=9$ is the number of hourly intervals correspond to daytime period from 9h to 18h. We have $\Phi_j = I^j - \Psi$ as the deviation of an individual's behavior from the mean, where I^j is behavior of subject j , Ψ is the average behavior of the society. Next, we construct the covariance matrix from the set of Φ_j ^[5]:

$$C = \frac{1}{H} \sum_{j=1}^H \Phi_j \Phi_j^T = AA^T \quad (1)$$

where $A = [\Phi_1, \Phi_2, \dots, \Phi_M]$. From this behavioral covariance matrix, we compute the set of eigenvectors (principal components) u_1, u_2, \dots, u_H , called eigenbehaviors, which can help to reconstruct individual's behavior by following transformation:

$$w_k^j = u_k (I^j - \Psi) \quad (2)$$

for $k = 1, 2, \dots, H$. In this PCA-based method, a reduced number h , where $h < H$, of eigenbehaviors that associated with h largest eigenvalues, could be sufficient for identifying subjects. The reconstruction weights vector representing subject j can be formed: $\Omega^j = [w_1^j, w_2^j, \dots, w_h^j]^T$. Similar to [7], the Euclidean distance between reconstruction weights vectors is utilized as the metric for describing similarity:

$$d(j, m) = \|\Omega^j - \Omega^m\| \quad (3)$$

$d(j, m)$ is the distance between subject j and m , Ω^j and Ω^m correspond to the reconstruction weights vectors of subject j and m .

1.2 Social relationship based similarity

In this method, we consider employing the social relationship factor to measure the similarity between subjects. Our motivation is that the relationship between person A and person B could be reflected through the relationship between A and the person who is socially closest B. In general, if A and B have high level of closeness, then A tends to be intimate with those who are socially close to B as well. Thus, to estimate social affinity between an individual and another subject in the society, the metric should include the factors which represent the social relationship between the individual and the socially closest persons of that subject.

Given an individual j in the society, to estimate the similarity between him and a subject m , the social relationships between j and those who are socially closest to m will be taken into account. Here, we have to determine the set of socially closest persons of a

subject. In our method, the degree of social relationship of two subjects will be reflected by w , the rate of contacts (the number of contact events for each day) between them. Let m_i be the i^{th} socially closest person of m . Among the subjects in the society, m_i is chosen as the one who has i^{th} highest rate of contacts with m . Then, to calculate the similarity between an individual j and subject m in the society, for general case, we include not only their direct social relationship estimation $w(j,m)$, but also the social relationship factors from the top K socially closest persons of m :

$$w_{SR}(j,m) = \alpha w(j,m) + \sum_{i=1}^k \beta_i w(j,m_i) \quad (4)$$

For instance, m_1 , the first socially closest person of m , is elected as $\arg \max_{h \neq j} w(m,h)$. The term $h \neq j$ means we exclude the current partner j from considering the person who is socially close m , to avoid including additional value 0 to similarity through element $w(j,j)$ (contact rate between j and himself is 0), in case j is the one who has highest contact rate with m in the society. Similarly, m_i is the one whose contact rate with m , $w(m,m_i)$ is ranked number i among the subjects in the society, excluding j . The sum of all weight parameters $\alpha + \sum_i \beta_i = 1$, in which α is the weight of direct social relationship (represented by contact rate) between j and current partner m , β_i specifies the weight of i^{th} social relationship factor (weight of contact rate between j and m_i) to the overall similarity value.

To make the value of this metric appropriate for clustering process, we normalize the similarity value. We adapt feature scaling to those similarities into the interval [0 1], and then, for the purpose of removing any existing asymmetric, we take the average value of $w_{SR}(j,m)$ and $w_{SR}(m,j)$ to represent the pairwise similarity between j and m . Finally we obtain the social relationship based similarity metric, where the larger value indicates the higher level of closeness.

2. Clustering social group

After measuring social closeness, we cluster mobile users into groups of strong social cohesion. Spectral clustering^[13] is an appropriate algorithm that can model the pairwise similarity or distance into relationship between subjects. In this paper, we perform spectral clustering and apply symmetric normalized technique^[14].

We first build an adjacency matrix, in which elements represent the local neighborhood similarities between subjects. Based on the adjacency matrix, the Laplacian graph matrix is constructed. Then, we map subjects to a lower-dimensional space, which is formed by the eigenvectors (that correspond to the smaller eigenvalues) of Laplacian matrix. Finally, we perform K-means algorithm on this data space and obtain the social groups.

IV. Evaluation Method and Results

1. Friendship-based evaluation method

The aim of the proposed scheme is to uncover the community structure of society, in which subjects in the same group should have close ties. In an effort to evaluate the clustering performance, we propose a new mechanism: friendship-based evaluation method. In general, there are three common properties of friendship: regularity, frequency and longevity^[8]. The social closeness of two subjects is demonstrated through how frequently, how regularly, and how long they are in contact with other. Therefore, those characteristics between subjects could be employed to estimate the level of closeness of relationship.

We start by quantifying frequency characteristic. Frequency can be represented by the rate of contacts between subjects. The frequency metric between two subjects is calculated as the ratio of the total number of contact events between them to the number of days of their overlapping period.

Regularity may be linked to the index of dispersion

^[12], the ratio of variance to the mean of the inter-meeting time intervals. Two subjects have regular relationship if the intervals between their meetings are equivalent, in other words, the variance of those time intervals is small. Therefore, the regularity between two subjects can be derived as the inverse of the index of dispersion of the inter-meeting time intervals.

Longevity is referred to the time length of the meeting events between subjects. If two subjects stay for a longer duration when they meet each other, then apparently they have closer relationship. The longevity metric is calculated as the average of meeting times between two subjects.

After determining relationship values for the society, we utilize them as the metrics to evaluate clustering scheme. Firstly, the asymmetric of the calculated metrics (e.g. the difference between pairwise value (m_i, j) and (j, m) , if exist) is removed. To make it easier for the observation, for each of three relationship metrics, we normalize the values to the interval $[0, 100]$. Finally, we compute the friendship metrics based on clustering result. Given a clustering scheme, if the society is partitioned to k social groups, where i and j is one pair of subjects that both belong to set of subjects in group n , then, for each relationship metric such as longevity, frequency or regularity, the overall performance R of clustering method is calculated as follows:

$$R = \frac{\sum_{n=1}^k r(i, j)}{N(i, j)} \quad (5)$$

where (i, j) represents a pair of subjects that belongs to the same cluster n , $r(i, j)$ corresponds to one of the three friendship metrics above, and accordingly, $N(i, j)$ is the number of pairs.

Through above steps, we derived the mechanism to estimate the overall frequency, regularity and longevity that illustrate the level of closeness between subjects in the society, thus can help to evaluate the clustering performance.

2. Results and Discussion

To validate our social analysis scheme, here we introduce another similarity measurement method, named contact count based. This simple metric estimates the pairwise social closeness between two subjects by counting the number of daily contacts between them, which is computed as the ratio of total number of contacts to the numbers of days in their overlapping period. For the social relationship based metric, in this paper, we select K equals to 1, that means we focus on employing in the factor of the first socially closest person. The weight parameter α is chosen as 0.5, and $\beta_1 = 0.5$.

Figure 3 shows the friendship-based evaluation results for the proposed schemes. Our schemes are compared with eigenbehavioral similarity^[7] based

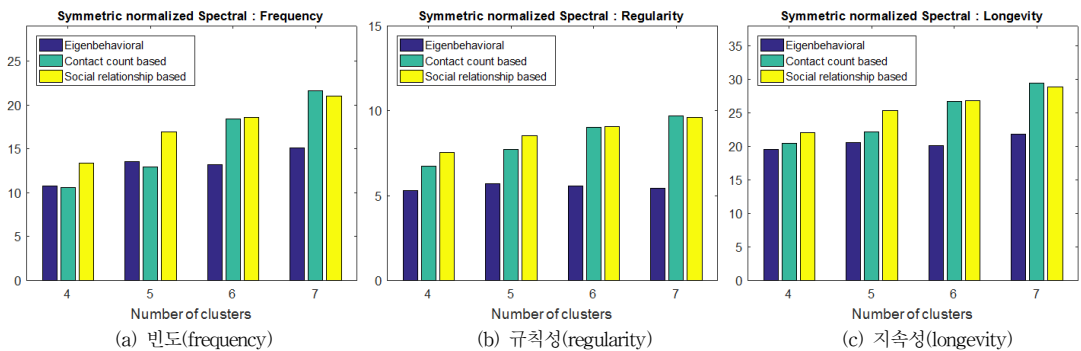


그림 3. 빈도, 규칙성 및 지속성에 따른 사회관계구조 검출 성능 평가

Fig. 3. Performance evaluation of social community detection based on frequency, regularity and longevity

scheme. We perform clustering for four cases, with the number of social groups, k , is chosen from 4 to 7, and collect the friendship metrics. In each case, following the formula (5), we derive the overall longevity, frequency and regularity by computing the means of relationship values between group mates, in which the social groups and their members are obtained through clustering.

As we can see in Fig. 3, contact count based and social relationship based are better than eigenbehavioral method. Eigenbehavioral similarity only considers the number of subjects that each individual was in contact rather than who really were in contact with him, whereas these two other methods embed that context.

In addition, as shown in Fig. 3, we can recognize the advantage of employing social relationship factor to the simple contact count based metric when k is small ($k = 4$ and 5); when k becomes larger, the performance of these two methods seems to be not so different. Accordingly, only employing the social relationship factor of one socially closest person gives restricted enhancement.

As can be observed, the contact count based and social relationship based similarity method show different performances. When the society is only partitioned into $k = 4$ or 5 clusters, the number of members in each group is large; thus there may exist some group members with limited relations and similar rate of contacts. In these cases, employing social relationship factor could give more information and help to better estimate relations between subjects than only counting the number of daily contacts. While with larger k , the size of clusters becomes smaller, the social-group-mates would have stronger cohesions that can be well-reflected directly through rate of contacts; thus including the social relationship factor is not so useful in compare with the simple contact count based method.

In general, three properties of the relationship show similar tendencies while evaluating clustering scheme. The larger number of clusters, the higher

community-level social closeness can be achieved. In other words, if the society is partitioned to larger number of groups, the overall friendship property values will increase.

V. Conclusion

In this paper, we have presented the methods that measure the human social closeness based on their contact histories. We have examined the communities within mobile users network by extracting human behavioral information. Through clustering subjects into numerous social groups, we aim to explore underlying community structure of a social network. We also proposed the evaluation method that employs friendship properties. The result suggests that separating the society into a large number of groups would increase the level of social ties and interdependence between subjects in the same community.

Based on proposed scheme, the communities of subjects with close ties and correlative social characteristics could be determined. This facilitates the opportunity of understanding relationships and interactions among mobile users, thus may lead to further applications, such as optimization for social-based opportunistic networks .etc.

References

- [1] Y. Zhu, B. Xu, X. Shi and Y. Wang, "A survey of social-based routing in delay tolerant networks: Positive and negative social effects," *IEEE Communications Surveys & Tutorials*, Vol. 14, no. 3, pp. 1–15, 2012.
DOI : <https://doi.org/10.1109/SURV.2012.032612.00004>
- [2] T. Minami, K. Baba, "An Attempt to Find Potential Group of Patrons from Library's Loan Records," *International Journal of Internet, Broadcasting and Communication*, Vol. 6, no. 1, pp.

- 5-8, Feb. 2014.
DOI : <https://doi.org/10.7236/IJIBC.2014.6.1.5>
- [3] D. Hwang, M.Y. Paek, "Differentiated impacts of SNSs on Participatory Social Capital in Korea," International Journal of Internet, Broadcasting and Communication, Vol. 8, no. 3, pp. 1-11, Aug. 2016.
DOI : <https://doi.org/10.7236/IJIBC.2016.8.3.1>
- [4] N. Eagle, A. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," Proc. Natl. Acad. Sci. USA, Vol. 106, no. 36, pp. 15274-15278, Sep. 2009.
DOI : <https://doi.org/10.1073/pnas.0900282106>
- [5] M. E. J. Newman, "Modularity and community structure in networks," Proc. Natl. Acad. Sci. USA, Vol. 103, no. 23, pp. 8577-8582, 2006.
DOI : <https://doi.org/10.7236/IJIBC.2005.5.2.56>
- [6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of community hierarchies in large networks," J. Stat. Mech., Vol. 008, no. 10, P10008, 2008.
DOI : <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- [7] N. Eagle, A. Pentland, "Eigenbehaviors: Identifying structure in routine," Behavioral Ecology and Sociobiology, Vol. 63, pp. 1057-1066, Sep. 2009.
DOI : <https://doi.org/10.1007/s00265-009-0739-0>
- [8] E. Bulut and B. Szymanski, "Exploiting friendship relations for efficient routing in delay tolerant mobile social networks," IEEE Trans. Parallel Distrib. Syst, Vol. 23, no. 12, pp. 2254-2265, 2012.
DOI : <https://doi.org/10.1109/TPDS.2012.83>
- [9] J.-P. Onnela et al., "Structure and tie strengths in mobile communication networks," Proc. Natl. Acad. Sci. USA, Vol. 104, no. 18, pp. 7332-7336, 2007.
DOI : <https://doi.org/10.1073/pnas.0610245104>
- [10] N. Eagle, A. Pentland, "Reality mining: sensing complex social systems," Journal Personal and Ubiquitous Computing, Vol. 10, issue 4, pp. 255-268, May. 2006.
DOI : <https://doi.org/10.1007/s00779-005-0046-3>
- [11] M. Turk and A. Pentland, "Eigenfaces for Recognition," J. Cognitive Neuroscience, Vol. 3, no. 1, pp. 71-86, 1991.
DOI : <https://doi.org/10.1162/jocn.1991.3.1.71>
- [12] D. R. Cox, P. A. W. Lewis, The Statistical Analysis of Series of Events, Chapman and Hall, London, 1966.
- [13] U. von Luxburg, "A tutorial on spectral clustering," Technical Report 149, Max Planck Institute for Biological Cybernetics, Aug. 2006.
- [14] Ng. Andrew, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," Advances in Neural Information Processing Systems, Vol. 14, MIT Press, pp. 849 - 856, 2002.

저자 소개

웬 풍 빈(준회원)



- August, 2014. B. Sc in Electrical Engineering, Hanoi University of Science and Technology, Vietnam.
 - Sep. 2015~Present. MS Student in Dept. of Electrical and Computer Engineering, University of Ulsan
- <Interest Areas : Wireless Sensor Network, Mobile Crowd Sensing>

윤 석 훈(정회원)



- 2000년 2월 인하대학교 자동화공학과 공학사
 - 2005년 6월 뉴욕주립대 (SUNY at Buffalo) 컴퓨터공학석사
 - 2009년 9월 뉴욕주립대 (SUNY at Buffalo) 컴퓨터공학박사
 - 2009~2011 LIG 넥스원 책임연구원
 - 2011~현재 울산대학교 IT융합학부 부교수
- <주관심분야 : 모바일네트워크, 무선센서네트워크, 최적화알고리즘>

※ 이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2016R1D1A3B03934617)