

A Concept of Applying Social Network Analysis in Medical Diagnosis

Sunantha Sodsee

Faculty of Information Technology
King Mongkut's University of Technology North
Bangkok
Bangkok, Thailand
sunanthas@kmutnb.ac.th

Maytiyanin Komkhao

Faculty of Science and Technology
Rajamangala University of Technology Phra Nakhon
Bangkok, Thailand
maytiyanin.k@rmutp.ac.th

Abstract—This paper aims to apply a social network analysis (SNA) to a medical diagnosis issue, which is the Cesarean sections due to Cephalopelvic disproportion (CPD). Firstly, the pregnant women, here is called “Patients”, are connected with each other by a medical examination consideration. The patients are also grouped based on medical pattern similarity. Secondly, the centrality measures in SNA, such as Degree, Hub or Authority, Closeness centrality, and Betweenness Centrality are applied to identify the cluster representative in each group of patients. To evaluate the proposed idea, the sample test of patients is conducted. The results show that hub or authority is taken to identify the cluster representative with minimal average distance between the test patient and existing groups, which is 0.684 by Euclidean distance. Finally, the medical diagnosis of test patient, here, is also acceptable by comparing to its real diagnosis by physician.

Keywords—social network analysis; centrality; clustering; medical diagnosis;

I. INTRODUCTION

At present, a social network analysis (SNA) has recently been an interesting research topic in several fields such as social science [1], computer science [2], and engineering [3]. As known, social network is a network of people or organization that connected with each other by social relationships such as friend, reputation, and co-author [4]. For its benefits, SNA also provide us to understand the behaviors or characteristics of relationship networks, which are analysed in graph-theoretical terms [5].

To deal with a medical issue, SNA has been applied for monitoring health of a group of people such as [6]. In addition, the influence of social networks, which are also important influences on health-related behaviors, has been studied for analysing risks of diseases, for instance, risk for HIV/AIDS [7].

As mention, in this paper, we present the influence of social networks in a medical diagnosis. The delivery of pregnant women [8], here, is considered as a medical issue. For diagnosing, the SNA is applied with clustering algorithms in order to give patients delivery prediction: Cesarean sections due to Cephalopelvic disproportion or the normal delivery.

This paper is organised as follows. Section II provides the overview of proposed method, applying SNA to understand the patient relationship, and description of medical examination dataset of patients. To evaluate this method, experimental analysis is given in Section III, and Section IV indicates a conclusion of this work.

II. DIAGNOSING DISEASE BY SOCIAL NETWORK ANALYSIS CONCEPT

This part explains detail of the proposed method: diagnosing disease by social network analysis.

A. Detail of Diagnosing Disease by SNA

In this work, medical examinations of patients, here, are considered to “*patient relationship*”, which is based on similarity of medical examinations. From Fig. 1, the patient relationship is created as a network. The network is clustered or grouped by the clustering algorithm. Each group presented as a disease contains similar patients in the case of medical examinations and “key” patient is selected acting as a group representative by SNA.

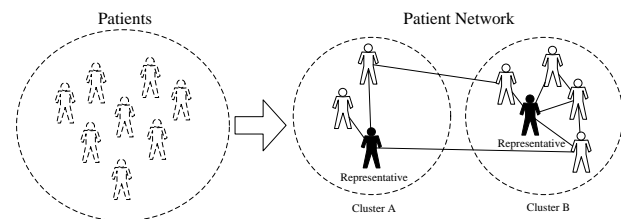


Figure 1. Identifying Cluster Representative by Social Network Analysis

To diagnose disease, the test patient is compared to the “key” patient (group representative) of each disease cluster.

The proposed method is described in detail as follows:

1. Creating network represented by graph
 - a. Defining network $G=(V,E)$. G is an undirected graph represented a patient relationship. Nodes are presented as medical examinations of patients, as well as links or edges are shown relationship between patients, which are connected by similarity

- identification of medical examinations compared to similarity threshold.
- b. Defining similarity threshold for connecting each patient as a similarity attachment condition.
- c. Creating an adjacency matrix represented the patient relationship.
- d. Updating network, new patient can be added to the existing network continuously by similarity attachment condition (see (b)).
- 2. Analysing the network from adjacency matrix
 - a. Calculating centrality scores of each patient by centrality measures: Degree Centrality, Hub or Authority, Closeness Centrality, and Betweenness Centrality [5].
- 3. Clustering the network by clustering algorithm
 - a. Defining the number of clusters.
 - b. Applying clustering algorithms such as k -Nearest Neighbor Analysis (k -NN) [9] to group the patients.
- 4. Applying the centrality scores to define the key patient as a representative of each cluster, this is also presented as a model of disease cluster.
- 5. Evaluating the proposed method by diagnosing disease to test patient, then compared to the real diagnosis by physician.

B. Identifying Key Patient: Representative of Group

The purpose of this sub section is to identify the key patient or representative in each group by social network analysis concept, which is a centrality measure.

Herein, we consider a patient relationship, which is represented by an undirected graph G . It is $G(V, E)$ of the set of patients, $V = \{v_1, v_2, v_3, \dots, v_N\}$, with $N = |V|$, and the set E of relationship between patients: $e_{i,j} = (v_i, v_j) \in E$ that means v_i and v_j are direct neighbours.

In fact, this undirected graph may be interpreted as a bi-directed graph of relationship. Due to its relationship is represented by non-directed arrows, which is also equivalent to bidirectional arrows connecting patients.

To characterise a key patient, we mainly consider Degree Centrality, Closeness Centrality, shortest-path Betweenness Centrality, Hub and Authority as follows [5].

First, Degree Centrality (DC) indicates the number of a patient's neighbours within one hop of connection. It is expressed as

$$DC(v_i) = \frac{d(v_i)}{N-1} \quad (1)$$

where $DC(v_i)$ represents the degree centrality of patient v_i , $d(v_i)$ is the degree of patient i , and N is the total number of patients in the network.

Secondly, Closeness Centrality (CC) describes the efficiency of medical examination propagation from one

patient to others. The closeness centrality of patient i is described as

$$CC(v_i) = \frac{1}{\sum_{i \neq j} \text{distance}(v_i, v_j)} \quad (2)$$

where $\text{distance}(v_i, v_j)$ is a measure for the distance between patient i and patient j .

Thirdly, shortest-path Betweenness Centrality (BC) indicates how often a patient occurs on the shortest paths between the indirectly connected patients. The betweenness centrality of patient i is given by

$$BC(v_i) = \sum_{i \neq j} \frac{sp_{j,k}(v_i)}{sp_{j,k}} \quad (3)$$

where, $sp_{j,k}(v_i)$ presents the number of shortest paths connecting patients j and k which pass through patient i , and $sp_{j,k}$ is the number of shortest paths connecting patients i and j .

Finally, the measure Hub describes patients that are connected to a large number of important ones, i.e. they are highest-degree patients. In contrast, Authority characterises the patients that connect many important ones. If G is a(n) undirected/bi-directed graph, then Hub is equivalent to Authority too.

C. Medical Examination of Patients

To present the performance in disease diagnosis, in this work, we concern about the delivery of pregnant women.

Here, the medical examination dataset of pregnant women, who delivered in a public hospital in Thailand between 1 July 2005 and 31 May 2007 [8], is considered randomly as a sample data for our work. It contains two groups of patients who delivered by Caesarean section due to cephalopelvic disproportion (called CPD class) and delivered normally (called NL class).

This dataset contains nine attributes selected by experts, which are Maternal Age, Nulliparous Parity, Gravidity, Pre-pregnancy Weight, Weight, Maternal Height, Pre-pregnancy BMI, BMI before Delivery and Fundal Height. In order to make different attributes comparable, they need to be normalised to a common scale by min-max normalisation, which is calculated by $\frac{(\text{value} - \min)}{(\max - \min)}$.

Finally, these medical examinations of patients are separated to training and test data. The proposed method aims to diagnose a test data, which is a test patient, to the Caesarean sections due to Cephalopelvic disproportion: CPD class, or the normal delivery: NL class.

III. SIMULATION RESULTS

In this section, the simulation results are presented as an evaluation of the proposed work. It contains three major parts of simulations, which are the analysis of studied network, cluster or group representative comparison, and medical diagnosis for test patient.

A. Network Analysis

For understanding, here, example of 20 patients is randomly selected as training data. Its network is constructed based on the proposed idea with the similarity threshold = 0.6 (see Section II). Table I presents the analysis of the studied network by using the social network analysis concept, for instance Degree Centrality, Hub or Authority, Closeness Centrality and Betweenness Centrality.

P-1 is the isolated node in this network then it has the lowest centrality scores, which are “0”. On the other hand, *P-20* is a big central node in the network because of its Degree, Closeness Centrality, and Betweenness Centrality scores.

Considering Hub or Authority of each patient, *P-4* has the highest score of it, which is 0.299. In contrast, Degree, Closeness Centrality, and Betweenness Centrality scores of *P-4* cannot reach to the maximal values, which are 12, 0.7125, and 0.0327, respectively.

TABLE I. EXAMPLES OF NETWORK ANALYSIS

| Patient | Degree | Hub/Authority | Closeness Centrality | Betweenness Centrality |
|-------------|--------|---------------|----------------------|------------------------|
| <i>P-1</i> | 0 | 0 | 0 | 0 |
| <i>P-2</i> | 10 | 0.261 | 0.633 | 0.0064 |
| <i>P-3</i> | 11 | 0.282 | 0.684 | 0.0138 |
| <i>P-4</i> | 12 | 0.299 | 0.7125 | 0.0327 |
| <i>P-5</i> | 8 | 0.167 | 0.6107 | 0.017 |
| <i>P-6</i> | 9 | 0.235 | 0.633 | 0.0072 |
| <i>P-7</i> | 9 | 0.24 | 0.6107 | 0.0021 |
| <i>P-8</i> | 9 | 0.226 | 0.6333 | 0.0162 |
| <i>P-9</i> | 5 | 0.096 | 0.5516 | 0.001 |
| <i>P-10</i> | 5 | 0.095 | 0.5029 | 0 |
| <i>P-11</i> | 11 | 0.282 | 0.684 | 0.0138 |
| <i>P-12</i> | 11 | 0.23 | 0.684 | 0.07 |
| <i>P-13</i> | 6 | 0.117 | 0.57 | 0.0021 |
| <i>P-14</i> | 8 | 0.141 | 0.6107 | 0.0136 |
| <i>P-15</i> | 10 | 0.188 | 0.6577 | 0.0401 |
| <i>P-16</i> | 11 | 0.278 | 0.684 | 0.0254 |
| <i>P-17</i> | 12 | 0.288 | 0.7125 | 0.0473 |
| <i>P-18</i> | 8 | 0.215 | 0.5897 | 0.0014 |
| <i>P-19</i> | 12 | 0.239 | 0.7125 | 0.0609 |
| <i>P-20</i> | 13 | 0.289 | 0.7435 | 0.102 |

B. Cluster Representative

All patients, here, are grouped by using clustering algorithm, such as *k*-Nearest Neighbor Analysis (*k*-NN), which is based on their medical examination pattern. Herein, there are four groups available, which are presented by “*G*₁”, “*G*₂”, “*G*₃”, and “*G*₄”. The members of each group are as follows:

$$G_1 = \{P-1, P-5, P-10\},$$

$$G_2 = \{P-2, P-3, P-6, P-7, P-11, P-12, P-16, P-18, P-20\},$$

$$G_3 = \{P-4, P-8, P-15, P-17\}, \text{ and}$$

$$G_4 = \{P-9, P-13, P-14, P-19\}.$$

To identify the key patient: representative of each cluster or group, the centrality measures: Degree, Hub or Authority, Closeness centrality, and Betweenness centrality, here, are used to consider it. The group representative is identified by the member with highest score of considered centrality measure compared to other members in the group.

1) Considering Degree for Group Representative

Table II shows the group representative of each group by Degree consideration. For *G*₁, *G*₂ and *G*₄, *P-5*, *P-20* and *P-19* are the representative with Degree score = 8, 13, and 12, respectively. On the other hand, *G*₃ has two representatives (*P-4* and *P-17*) because of their equal Degree values.

TABLE II. IDENTIFYING REPRESENTATIVE BY DEGREE CONSIDERATION

| Group | Representative | Class |
|-----------------------|------------------|-------|
| <i>G</i> ₁ | <i>P-5</i> | CPD |
| <i>G</i> ₂ | <i>P-20</i> | NL |
| <i>G</i> ₃ | <i>P-4, P-17</i> | CPD |
| <i>G</i> ₄ | <i>P-19</i> | CPD |

2) Considering Hub or Authority for Group Representative

Each group: *G*₁, *G*₂, *G*₃, and *G*₄, has *P-5*, *P-20*, *P-4*, and *P-19* being the representative of group with Hub or Authority score = 0.167, 0.289, 0.299, and 0.239, respectively (see Table III).

TABLE III. IDENTIFYING REPRESENTATIVE BY HUB/AUTHORITY CONSIDERATION

| Group | Representative | Class |
|-----------------------|----------------|-------|
| <i>G</i> ₁ | <i>P-5</i> | CPD |
| <i>G</i> ₂ | <i>P-20</i> | NL |
| <i>G</i> ₃ | <i>P-4</i> | CPD |
| <i>G</i> ₄ | <i>P-19</i> | CPD |

3) Considering Closeness Centrality for Group Representative

Each group has *P-5*, *P-20*, *P-4* and *P-17*, and *P-19* being the representative of group with Closeness centrality score = 0.6107, 0.7435, 0.7125, and 0.0609, respectively (see Table IV), which are the same results as the degree consideration.

TABLE IV. IDENTIFYING REPRESENTATIVE BY CLOSNESS CENTRALITY CONSIDERATION

| Group | Representative | Class |
|-----------------------|------------------|-------|
| <i>G</i> ₁ | <i>P-5</i> | CPD |
| <i>G</i> ₂ | <i>P-20</i> | NL |
| <i>G</i> ₃ | <i>P-4, P-17</i> | CPD |
| <i>G</i> ₄ | <i>P-19</i> | CPD |

4) Considering Betweenness Centrality for Group Representative

Table V presents the group representative of each group by considering Betweenness centrality. For G_1 , G_2 , G_3 and G_4 , $P-5$, $P-20$, $P-17$ and $P-19$ are the representative with Betweenness centrality score = 0.017, 0.102, 0.0473 and 0.0609, respectively.

TABLE V. IDENTIFYING REPRESENTATIVE BY BETWEENNESS CENTRALITY CONSIDERATION

| Group | Representative | Class |
|-------|----------------|-------|
| G_1 | $P-5$ | CPD |
| G_2 | $P-20$ | NL |
| G_3 | $P-17$ | CPD |
| G_4 | $P-19$ | CPD |

C. Medical Diagnosis for Test Patient

Herein, the medical diagnosis of a test patient is focused. This test patient data is selected randomly from the medical examination dataset of pregnant women. The result of diagnosis, here, is determined by minimal measured distance between the test patient data and the representative of groups with Euclidean distance or City Block distance.

The result shows that the test patient data closes up to G_4 (for all considered centrality measures) with minimal measured distances, which are 0.499 by Euclidean distance and 1.207 by City Block distance. It means that the diagnosis result of test patient will be “CPD”: Caesarean sections due to Cephalopelvic disproportion. This result is evaluated by comparing to the real diagnosis from physician, which is “CPD”. The result shows that this diagnosis is correct.

Table VI presents an example result of test patient diagnosis when hub or authority is considered. Its nearest group is G_4 showing that the diagnosis of test patient is “CPD”.

TABLE VI. PREDICTING RESULT WHEN HUB OR AUTHORITY CONSIDERED

| Group | Euclidean | | City Block | |
|-------|-----------|-------|------------|-------|
| | Distance | Class | Distance | Class |
| G_1 | 0.756 | CPD | 2.046 | CPD |
| G_2 | 0.702 | NL | 1.659 | NL |
| G_3 | 0.782 | CPD | 1.73 | CPD |
| G_4 | 0.499 | CPD | 1.207 | CPD |

TABLE VII. AVERAGE DISTANCE MEASURED IN EACH CENTRALITY

| Centrality Measure | Euclidean | City Block |
|--------------------|-----------|------------|
| Degree | 0.6915 | 1.694 |
| Hub/Authority | 0.685 | 1.660 |
| Closeness | 0.6915 | 1.694 |
| Betweenness | 0.712 | 1.857 |

Finally, the average distance measured between the test patient data and group representatives are considered in Table VII. For the minimal average distance, it shows that

the test patient is more similar to patients in existing groups, when hub or authority is considered to identify the group representatives than other measures.

IV. CONCLUSION

This paper presents an applying social network analysis (SNA) in Caesarean sections due to Cephalopelvic disproportion diagnosis. The patient medical examinations are grouped by clustering algorithms such as k -Nearest Neighbor Analysis; the centrality measures: degree, hub or authority, Closeness centrality, and Betweenness centrality from SNA are, then, used to identify the representative of each group. To diagnose the test patient, Euclidean and City block distances are applied to measure similarity among existing representative of medical examination groups. The result shows that the medical diagnosis of test patient is correct. It is evaluated by comparing to the real diagnosis.

For further work, large medical dataset of pregnant women will be considered in order to measure *sensitivity*, *specification*, and *accuracy* of diagnosis.

ACKNOWLEDGMENT

This paper was supported by Faculty of Information Technology, King Mongkut's University of Technology North Bangkok and Faculty of Science and Technology Rajamangala University of Technology Phra Nakhon, Thailand.

REFERENCES

- [1] C. Lin, L. Wu, Z. Wen, H. Tong, V. Griffiths-Fisher, L. Shi, and D. Lubensky, “Social Network Analysis in Enterprise,” *Proceedings of the IEEE*, vol.10, iss. 9, pp. 2759 – 2776, 2012.
- [2] F. Antunes and J. Paulo Costa, “Integrating Decision Support and Social Network,” *Advances in Human-Computer Interaction*, vol.2012, January 2012.
- [3] E. Stai, V. Karyotis, and S. Papavassiliou, “Exploiting Socio-Physical Network Interactions via a Utility-Based Framework for Resource Management in Mobile Social Networks,” *IEEE Wireless Communications Magazine*, vol.21, iss.1, pp.10-17, 2014.
- [4] E.F. Churchill and C.A. Halverson, “Social Networks and Social Networking,” *IEEE Internet Computing*, vol.9, iss.5, pp. 14-19, 2005.
- [5] S. P. Borgatti, M. G. Everett, “A Graph-Theoretic Perspective on Centrality,” *Social Networks*, vol.28, pp.466–484, 2006.
- [6] S. B. Gesell, S. L. Barkin, and T. W. Valente, “Social Network Diagnostics: a Tool for Monitoring Group Interventions,” *Implementation Science*, vol.8, 2013.
- [7] M.E. Morris, “Network Epidemiology: A Handbook for Survey Design and Data Collection,” New York, Oxford University Press, 2004.
- [8] A. Komkhao, “Delivery due to Cephalopelvic Disproportion in BhumibolAdulyadej Hospital, Thailand,” *Royal Thai Air Force Medical Gazette*, vol.54, iss.70, 2008.
- [9] J. Han, M. Kamber, and J. Pai “Data Mining: Concepts and Techniques,” 3rd ed. Waltham, MA, Morgan Kaufmann Publishers, 2012.