IJCSE International Journal of Computer Science and Engineering Open Access
Research Paper Volume-2, Issue-5 E-ISSN: 2347-2693

Social Networking Analysis

Shalki Chahar[1]

Amity University, Uttar Pradesh
[1]shalki.chahar@gmail.com

www.ijcseonline.org

Received: 19/04/2014 Revised: 26 /04/2014 Accepted: 13/05/ 2014 Published: 31 /05/2014

Abstract- In today's day and age, a rapid proliferation of technology has enabled efficient global communication. As a result, the last decade has seen social networking emerge as the backbone of global interactions. At the kernel of this advent, lies the concept of networks. Networks are arrangements of interconnections among a variety of entities. From this we can deduce that social networks are social structures comprising individuals and the interactions they have with each other. The computational analysis of these networks is known as social network analysis.

Keywords: *Social Network Analysis, Directed Acyclic Graph, Statistical Relational Learning*

1. INTRODUCTION

1.1. Overview

Social network analysis views the complex set of relationships existing between various members of any social strata as graphs. The nodes of these graphs represent people or entities and the directed edges or links connecting the nodes represent the interactions between these entities in a social context.

Social networks are immensely dynamic in behaviour, constantly growing and unpredictably changing with time due to the formation of new links between interacting entities. This gives rise to what is known as the link prediction problem. Accurately predicting the manner in which new links may develop in a certain interval of time is one of the major concerns of this analysis.

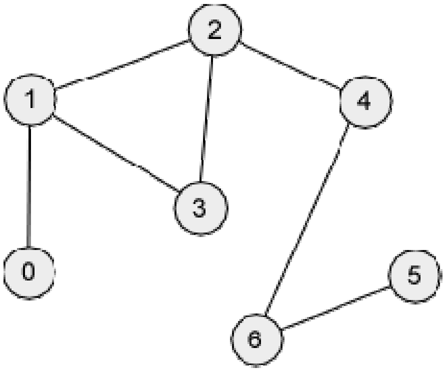
Another major area of study in this analysis is the hierarchical model. In this model, the position of any individual on the social network directly corresponds to their position in the social hierarchy. Therefore, with the help of this hierarchical organization missing links can be predicted from partially known social networks.

1.1 Theoretical Background

To adequately understand the concepts of social network analysis one has to understand the theories of structure and behaviour of networks. Due to their representation simplicity and clarity, Graph theory and Game theory are the key tools to comprehend networks and their mechanisms.

1.2.1. Graph Theory

Graph theory is the study of graphs. They consist of 'nodes' or 'vertices' and edges connecting these nodes. Graphs may be undirected or directed. In undirected graphs no distinction exists between the two vertices associated with each edge. On the other hand, directed graphs consist of vertices with their edges directed from one vertex to another.[1]



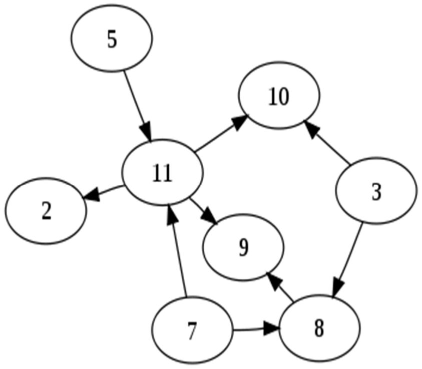


Fig1. Undirected graph
Fig 2. Directed graph

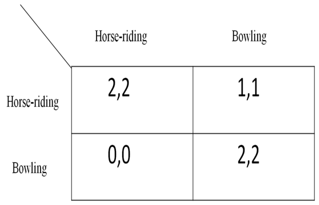
Graphs can be used to represent any relationship in physical, biological and social systems. A wide variety of practical processes can therefore be structured as graphs.

1.2.2. Game Theory

Classical game theory predicts how rational agents behave in strategies. Individuals are self-centred and optimize their own personal motives. Game theory therefore suggests that the joint outcome is directly affected by how a group of people in a social context decide to act individually. For example, let's consider a social scenario in which two friends, Opetha and Zeppelina who haven't contacted each

Corresponding Author: *Shalki Chahar*

other before committing to attending either one of the two activities: horse-riding or bowling. Opetha prefers horse-riding, while Zeppelina prefers bowling. However, both want to attend the same activity.[1] In that case, there are four possibilities as represented by the pay-off matrix:



	Horse-riding	Bowling
Horse-riding	2,2	1,1
Bowling	0,0	2,2

Fig 3. Pay off matrix

Thus we can conclude that game theory and graph theory are eminent tools to model networks and understand the static and dynamic properties of these networks.

1.3. Structural and Behavioral Connectivity Of Social Networks

Social networks encapsulate all types of interactions that exist between a group of people. These interactions may be of positive nature (friendship, love, support, respect, etc) or negative nature (hatred, contempt, harassment, etc). The two basic aspects which define these networks are: structural connectivity and behavioural connectivity. Structural connectivity implies whether a connection between two members represented by nodes exists or not. On the other hand, behavioural connectivity implies a more complex mechanism. The behavioural aspect has to be considered from both the network and the individual's perspective. Individual actions influence network characteristics and formation and vice versa.[2]

Real time social networks are constantly modified and highly unpredictable. Thus one of the major areas of study in social network analysis is to construct models which can accurately predict the formation and evolution of future social networks. As mentioned earlier, at the heart of this study lies the basic concept of links and their development thus giving rise to the issue of Link Prediction.

2. Hierarchical Structure

Social networks as seen are complex structures with various categories of study, the hierarchical structure in networks has been realised and studied. At the crust of hierarchy, lies the concept of social stratification. Since the beginning of human society, social stratification has been prominent dude to the existence of power, wealth, laws, etc.

Real world relations are characterized by two major features:

- They are a combination of both positive(supporting) and negative(opposing) interactions
- They are continuously evolving with time

Rise of social networks and their usage has facilitated the study of hierarchy existing between these complex networks which represent real world relations. Given a social graph, the social strata that an individual belongs to, may not be directly observed. However, nodes and links can be studied to infer hierarchy.

Given two nodes, 'u' and 'v' and there exist a link u →v which indicates that u is directed to v. If there is no reverse link from v to u, it might imply that v is higher up in the hierarchical structure than u. Various algorithms have been presented to evaluate the rank of nodes in a given network.

One of the most common and simple method of arranging the various nodes and link is through a tree. Hierarchy thus can be viewed as a partially ordered set, where each element of the set is a node and the partial ordering indicated the link from u to v. The graphs under study are directed graphs known as Directed Acyclic Graphs (DAGs).[4]

Consider a network, G = (V, E) where a rank r(v) is associated with every node, v. In social networks, individuals are aware of their position in the hierarchical structure. Therefore it can be implied that individuals having higher ranks will not connect to individuals having lower ranks. Hence directed edges will form from lower ranked nodes to higher ranked nodes. When the reverse occurs, i.e. higher ranked individuals connect to lower ranked individuals social agony is caused to these lower ranked nodes.[5]

In particular, whenever the rank of 'u' is less than the rank of 'v' ($r(u) < r(v)$) , a link from u to v (u →v) is expected and doesn't no cause any agony to u. In this case however if link v→u forms then this causes agony to 'v'. The amount of agony caused by each such reverse edge depends on the difference between their ranks and is equal to r(v)-r(u)+1. Hence the total agony in the network can be computed as:

$$A(G,r)= \sum_{(u,v)\epsilon E} \max(r(v) - r(u) + 1)$$

Hierarchy in a network can be inferred from the computed agony in the entire social network. Hierarchy in a directed graph G is defined as:

$$h(G)= 1- A(G)/m$$

$$= max_{r\epsilon Rankings}(1-\frac{1}{m}\sum_{(u,v)\epsilon E} \max(r(v) - r(u) + 1)$$

The value of hierarchy for any graph G always lies in [0, 1]. For the computation of hierarchy it is essential to evaluate the ranks of every node in a network representing real world relations. This can be done in a variety of ways. In this paper, two such ways are presented.

2.1. Hierarchy Status Evaluation Algorithms

2.1.1. Page Rank

Page Rank is an algorithm that efficiently computes the relative importance of each node in a graph by assigning a numerical weighting to every node. As seen earlier in the paper, page rank measure is a distribution that measures the probability of a random walk through the edges of the graph will arrive at a particular node.

By this algorithm the rank of a particular node 'u' can be calculated as :

$$R_u=\sum_{i\to u} \beta \frac{ri}{di}+(1-\beta)\frac{1}{n}$$

Where *di* denotes the out-degree of node 'u' and β denotes the probability of jumping to a random node.[6]

2.1.2. Number of Supporters

This algorithm computes the number of nodes in a network that can reach the node ‘n’ by following forward positive edges or backward negative edges. This measured value is known as the support number of each node. The status of the node is then ranked by support number of each node.
Therefore an adjacency matrix A∈$R^{n \times n}$ can be used to characterize a signed graph G as follows: A(i,j)=1 if there exists a link between the nodes i and j else A(i,j) =0.[7]

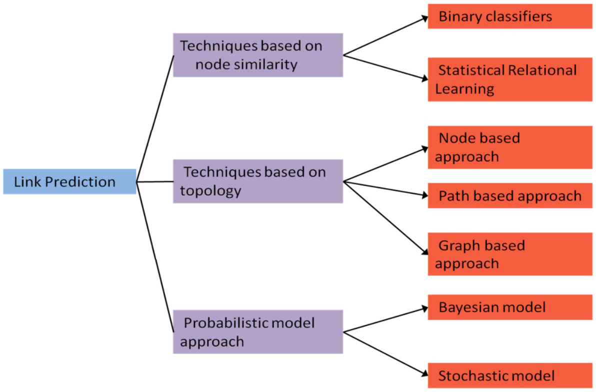
3. LINK PREDICTION

Research and study in the field of social network analysis have directly related the growth of networks to mathematical models. The link prediction problem focuses on four main principles:

- Link existence: determining whether a new link between two individuals will form in a future interval of time or not.
- Link type: determining the degree of relationship between two nodes of the network.
- Link cardinality: Determining the number of links connected to two nodes of the network.

All these factors play the primary role in the techniques and models presented for predicting the formation of links. These techniques can be categorised according to the approach of analysis into three divisions:

- Techniques based on node similarity
- Techniques based on topological arrangements
- Techniques based on probabilistic measures


Link Prediction
Techniques based on node similarity
Techniques based on topology
Probabilistic model approach
Binary classifiers
Statistical Relational Learning
Node based approach
Path based approach
Graph based approach
Bayesian model
Stochastic model

3.1. Techniques Based On Node Similarity

Node similarity models focus on the similarity existing between a pair of nodes to determine whether they will be linked in the future. Similarities are measured using machine learning techniques.[13][8]

Consider a social network S=(N,E) where N represents the set of nodes and E represents the set of edges existing between these nodes. The basic goal of link prediction is to determine whether a new edge e_{ij} will develop between a pair of nodes (n_i, n_j) in the network.

Lin in 1998 presented a technique to compute similarity between two nodes of a network. Similarity between two nodes (v_i,v_j) can be expressed as the ratio of the intersecting properties of the two nodes to the total set of properties related to each node.

$$sim(vi, vj) = \frac{\log P(common(Vi; Vj))}{\log P(description(Vi; Vj))}$$

Where Vi and Vj are the characteristics of nodes v_i and v_j.

3.1.1 Binary Classifiers

Binary classifiers evaluate the similar characteristic between two nodes based on their content information. This method is best suited for networks in which a pair of nodes has a number of similar properties. Mapping feature functions excerpt content features into a single vector â(v_i; v_j).For a node pair the link is reduced to :

$$link(vi; vj) \\ = \begin{cases} Does\ exist & if\ w\ \hat{a}(vi; vj) \\ Does\ not\ exist & if\ w\ \hat{a}(vi; vj) \end{cases}$$

3.1.2 Statistical Relational Learning

Statistical Relational learning deals with networks with very high uncertainty like real world social networking. In this methodology, link and node types are first determined. Taking in to consideration the entire array of features, nodes are classified into groups based on similarities. Furthermore, new links are induced on the basis of existing ones. SRL may be subdivided into directed and undirected approach. Directed approach focuses on causal interactions whereas undirected approach focuses on symmetric, non-causal interactions. SRL uses first order logic and represents uncertainty in the form of probabilistic graphs.[9]

3.2 Techniques Based On Topology

Topology based techniques recognize patterns or repetitive arrangements in a network to determine future links. Topological methods can be classified into three categories namely, node based approach, path based approach and graph based approach.

3.2.1 Node Based Approach

In this approach, the vicinity information of a node is taken into consideration. This includes evaluating the type and number of neighbouring nodes. The probability of two nodes of a network to link together is directly proportional to the number of common neighbours existing between the selected node pair.[10]
Various evaluation procedures have already been standardised by researchers. These procedures formulate a scoring function for a potential link that may develop between two existing nodes (v_i; v_j) of the network based on the number of neighbouring nodes Γ (v_i) and Γ (v_j) respectively.

- **Common Neighbors Method**

Newman in 2001 defined the scoring function of a probable link between two nodes as the number of common neighbours existing between the two nodes.

$$score(vi; vj) = |\Gamma(vi) \cap \Gamma(vj)|$$

- **Jaccard Coefficient**

This approach presents the scoring function as the ratio of the common neighbours between two nodes to the total number of their neighbours.

$$score(vi;vj) = \frac{|\Gamma(vi) \cap \Gamma(vj)|}{|\Gamma(vi) \cup \Gamma(vj)|}$$

- **ADAMIC/ADAR Coefficient**

Adamic and Adar in 2003 presented a similar scoring function, however they adopted a more accurate methodology. Firstly, similarity between nodes v_i and v_j is first computed as:

$$sim(vi;vj) = \sum_{z:feature\ shared\ by\ vi\ and\ vj} \frac{1}{\log(frequency(z))}$$

This modifies the other evaluation techniques by giving the uncommon features more weight rather than just counting of the common features. Secondly, the scoring function is represented as:

$$score(vi;vj) = \sum_{z \in (\Gamma(vi) \cap \Gamma(vj))} \frac{1}{\log|\Gamma(z)|}$$

Where $\Gamma(z)$ is the number of common features or nodes.

- **Preferential Attachment**

This model of link prediction has received considerable attention in determining the growth of social networks. This analysis evaluates the probability that a new link will have node v_i as the terminating point is proportional to the number of existing neighbours of v_i, $\Gamma(vi)$.Therefore, the probability that node v_i will receive a link from node v_j is proportional to the number of current neighbouring nodes of v_j and vice versa. Thus, the scoring function is defined as,

$$score(vi;vj) = |\Gamma(vi)|\ |\Gamma(vj)|$$

3.2.2 Path Based Approach

Path based approach primarily takes into consideration the shortest path or distance between two nodes. When two nodes are indirectly connected via a number of different links, the probability that a direct link will emerge between them is very high. Similar to the node similarity approach, various measures to evaluate path similarity have also been presented.[12] The primary techniques are as follows:

- **Katz Measure**

Katz measure proposed a scoring function for the link between two nodes as the sum of total number of paths existing between two nodes weighted according to their lengths. If $paths^{(l)}_{vi,vj}$ is the total number of paths existing between two nodes(v_i, v_j), then Katz measure may be formulated as,

$$score(vi;vj) = \sum_{l=1}^{\infty} \alpha^l\, paths^{(l)}_{vi,vj}$$

where α^l >0 is a parameter of the predictor.

- **Hitting Time Measure**

Hitting time measure defines a scoring function depending on the number of steps required to reach from one node, v_i to another node v_j assuming that a random walk is selected. The number of steps adopted is known as hitting time and is denoted by $H_{vi;vj}$. Hitting time is often not symmetric due to the assumption made, therefore a new term known as commute time, $C_{vi;\ vj}$ is evaluated.[11]

$$C_{vi,vj}=H_{vi,vj} + H_{vi,vj}$$

The scoring function is then evaluated by negating either the hitting time or the commute time.

- **Page Rank Measure**

This analysis technique is a modification of the hitting time mechanism. In this method, the scoring function of the link between two nodes v_i and v_j as the probability that node v_j appears on a random walk that is returning to node v_j. In this analysis a parameter, β ∈ [0,1] is defined as the probability of vj appearing in the arbitrary walk.

- **SIM Rank Measure**

This method of analysis presents a scoring function of the link between two nodes vi,vj on the basis of whether a similarity exists between the considered nodes and their neighbouring nodes.

$$score(vi,vj) = \frac{\beta\left(\sum_{x\epsilon\Gamma(vi)} \quad \sum_{y\epsilon\Gamma(vj)} score(x,y)\right)}{|\Gamma(vi)||\Gamma(vj)|}$$

4. PROBABILISTIC MODEL APPROACH

As the name suggests, this model of approach is based on simple laws of probability. They differ from the classical graph models, proposing a set of three models:

1. Data graph (Gd = (Vd Ed))
2. Model graph (Gm = (Vm Em))
3. Interference graph (Gi = (Vi Ei))

The key feature of this approach is that it reduces the link prediction problem to predicting the existence of similar attributes between two existing nodes for the formation of potential edges between them. In essence, this requires formulating an <exist> attribute.

The data graph (Gd = (Vd Ed)) contains a set of nodes, vi ϵ Vd and set of links between these nodes ei ϵ Ed. Each node and link has associated with them a type t, and a set of parameters corresponding to this type, Zt1. The joint probability distribution in the network thus cann computed as:

$$z= \{z^{tvi}_{vi} : vi \ \epsilon\, Vd ,\ T(vi)=t_{vi}\} U \{z^{tej}_{ej} : ej\ \epsilon\ Ed, T(ej)=t_{ej}\}$$

The model graph (Gm = (VmEm)) on the other hand categorises network entities with the same type. The technique formulates the the probabilistic dependencies between parameters of the same type. Other than the structure of interdependencies, the model graph also introduces conditional probability distributions (CPD) linked with the network nodes.

The interference graph (Gi = (ViEi)) is plotted based on the data and model graphs by instantiating sequence models. For each node-parameter pair in Gd, a local copy of the corresponding CPD from Gm is made in Gi.

- **Relational Bayesian Network(Rbn)**

RBNs are directed acyclic graphs (DAGs) represented by conditional probability distributions (CPDs) indicating the joint distribution over the network. A CPD corresponding to a parameter Z is specified by the probability, P (Z |pa(Z)|) where pa(Z) is the value of the parents of Z.

A network object is characterized by a set of parameters (Z1,Z2,.,Zn), the DAG representing the Bayesian Model is :

$$P(Z1,Z2,...,Zn) = \prod_{i=1}^{n} P(Zi\ |pa(Zi))$$

RBNs use closed form parameter estimation techniques providing a simple, accurate and efficient analysis method.

5. CONCLUSION

In this dissertation, link prediction and evaluation of hierarchy in real world social networks have been analysed. Graph and game theory basics have also been introduced to aid in this analysis.

Techniques based on node similarity, topology and probability have been efficient in the prediction of link formation in continuously evolving social networks. Various such algorithms have been clearly presented in this paper.

On the other hand, hierarchy in social networks has been computed by simply analysing the ranks of every individual represented by nodes in a social graph. There exists immense scope in the development of link prediction techniques in the field of social network analysis. For example, the effect of link strengths on the computation of link prediction can be considered. Overall, the foundation for many future works in this field has been laid.

REFRENCES

[1] Hierarchical structure and the prediction of missing links in networks by Aaron Clauset, Christopher Moore, M. E. J. Newman3.

[2] Predicting Hierarchical Structure in Small World Social Networks by Hussain, Dil Muhammed Akbar.

[3] Link Prediction In Social Networks by Prof. William H. Hsu.

[4] Networks, Crowds, and Markets: Reasoning about a Highly Connected World by David Easley and Jon Kleinberg. Cambridge University Press, 2010.

[5] The Link-Prediction Problem for Social Networks by David Liben-Nowell and Jon Kleinberg.

[6] Adamic, L.A., & Adar, E. (2003). Friends and neighbours on the Web. Social Networks, 25(3), 211–230.

[7] Statistical Relational Learning: A Tutorial by Lise Getoor University of Maryland, College Park ECML/PKDD 2007 Tutorial.

[8] Link prediction problem for soc net by David Liben-Nowell n Jon Kleinberg.

[9] Clauset, A., Moore, C., and Newman, M. E. J. Hierarchical structure and the prediction of missing links in networks.

[10] Citation: Social Network Analytics by Elena Pupazan.

[11] Emergence of global status hierarchy in social networks by Yue Chen, Jia Ji, Yizheng Liao.

[12] Ahmed, Elmagarmid, and Ipeirotis, Panagiotis G., and Verykios, Vassilios. (2007) Duplicate Record Detection: A Survey.

[13] Link prediction in social network by Muhammad Al Hassan n Muhammad J Zaki.