International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 3, Issue 2, Mar 2014, 1-8

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

SURVEY PAPER ON THE USAGE OF GRAPH MINING METHODS IN SOCIAL NETWORKING

JIGNESH PATEL[1], NIRAV SUTHAR[2] & BHAVESH OZA[3]

[1,2]M.E. Student, Department of Computer Science and Engineering, LD Engineering College, Ahmadabad, Gujarat, India

[3]Professor, LD Engineering College, Ahmadabad, Gujarat, India

ABSTRACT

This survey studies basic concepts of graph mining as well as Social Networks. Social networks have been widely used now-a-days such as Facebook, Linked-In, Google+, etc. Users of these sites form a social network, which provides a powerful means of sharing, organizing, finding contents and contacts. Social Network can be cast as graph. Users represented as "nodes" and their relationship is represented by "links". This allows us to characterize the network and analyze the network. Here presented some challenges in crawling.

KEYWORDS: Apriori Based Approach, Frequent Pattern Growth, Graph Mining, Graph Mining Approach, Social Network

INTRODUCTION

Social Network Analysis (SNA) is the study of relations between individuals including the analysis of social structures, social position, role analysis, and many others [29]. These are represented as graph with node representing individual or group and edges as relation between them.

With the prosperity of Internet and Web 2.0, many social networking and social media sites are emerging, and people can easily connect to each other in the cyber space. This also facilitates SNA to a much larger scale — millions of actors or even more in a network; Examples include email communication networks[1], instant messenger networks [2], mobile call networks [3]. Other forms of complex network like, biological networks, metabolic pathways, genetic regulatory networks, food web and neural networks, are also examined and demonstrate similar patterns [4]. So the analysis of social networks has recently experienced a surge of interest researchers. OSN data analysis has a great potential for researchers in a diversity of disciplines. However, we propose that OSN analysis should be placed in the context of its sociological origins and its basis in graph theory.

Graph Mining of on-line social networks is a relatively new area of research which however has a solid base in classic graph theory, computational cost considerations, and sociological concepts such how individuals interrelate, group together and follow one another [5]. The structure of this paper is as follow: section – 2 gives brief introduction about graphs and Social network. Section–3 gives introduction to graph mining. Section -4 and 5 gives challenges in crawling social network and crawling in social network respectively.

1. BASE SURVEY

2.1 Graphs

A graph *G* is represented as *G* (*V*, *E*) where *V* is a set of vertices (or nodes) and *E* is a set of edges (or links)

connecting some vertex pairs in *V*. Statistically, a graph can be characterized by derived values such as the average degree of the nodes and the average path length between nodes. Additional characteristics are the graphs diameter, the number of triangles, the number of is omorphisms and the clustering coefficient, among others.

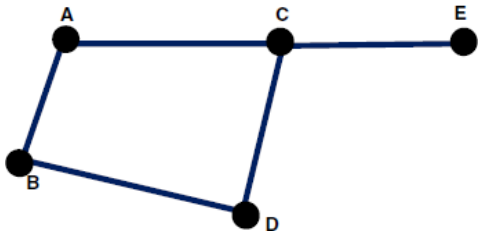


Figure 1: Simple Graph with Five Vertices and Five Edges

In figure 1 we see an elementary graph with five vertices and five edges. As there are no arrows, we assume it is undirected, and as the edges have no additional information attached we assume it is un-weighted. We see that nodes *A*, *B* and *D* have degree 2, node *C* has degree 3 and node *E* has degree 1, hence the degree sequence is {1, 2, 2, 2, and 3}.

There are many types of graphs: directed, undirected, graphs with weights on the edges, vertices or both. An 'undirected graph' has no information about the direction or flow between nodes. That is, the edge between two vertices *A* and *B* is identical to the edge between vertices *B* and *A*. A 'directed graph', on the other hand, *does* include directional information. Each edge will have a direction associated with it, which can be unidirectional $A \rightarrow B$ or bidirectional $A \leftrightarrow B$. A 'weighted graph' includes additional information associated with an edge or a vertex.

Graphs Used in Following Areas

- Internet/Computer Networks
- WWW
- Social Networks
- Transport Networks
- Many more …

2.2 Social Network

A Social Network is a social structure made up of a set of individual (or organization etc.) tied together by link. This link can be undirected or directed. Each individual is called an Actor and link is called the relationship between those actors. Both links and node have attributes.

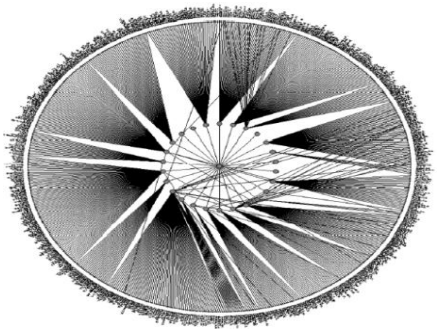


Figure 2: Science Co-Author Graph

2.2.1 Characteristics of Network

Almost all large real-world networks evolve over time by the addition and deletion of nodes and edges. Most of the recent models of network evolution capture the growth process in a way that incorporates two pieces of "conventional wisdom:"

- **Constant Average Degree Assumption:** The average node degree in the network remains constant over time. (Or equivalently, the number of edges grows linearly in the number of nodes.)
- **Slowly Growing Diameter Assumption:** The diameter is a slowly growing function of the network size, as in "small world" graphs [6-9].

But we observed following phenomena on some datasets …

- **Empirical Observation: Densification Power Laws.** The networks are becoming denser over time with the average degree increasing (and hence with the number of edges growing super linearly in the number of nodes). Moreover, the densification follows a power-law pattern.

$e(t) \propto n(t)^{a}$

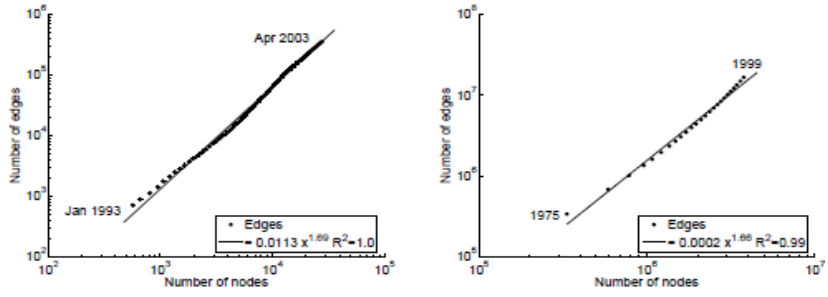
Where e(t) is number of edges and n(t) is number of nodes a is exponent strictly between 1 and 2

- **Empirical Observation: Shrinking Diameters**. The effective diameter is, in many cases, actually decreasing as the network grows [10].

We view the second of these findings as particularly surprising: Rather debating over exactly how the graph diameter grows as a function of the number of nodes.

Densification Law

Here we describe the datasets we used, and our findings related to densification. For each graph dataset, we have, or can generate, several time snapshots, for which we study the number of nodes $n(t)$ and the number of edges $e(t)$ at each timestamp t. We denote by n and e the final number of nodes and edges. We use the term *Densification Power Law plot* (or just DPL plot) to refer to the log-log plot of number of edges $e(t)$ versus number of nodes $n(t)$.



(a) **(b)**

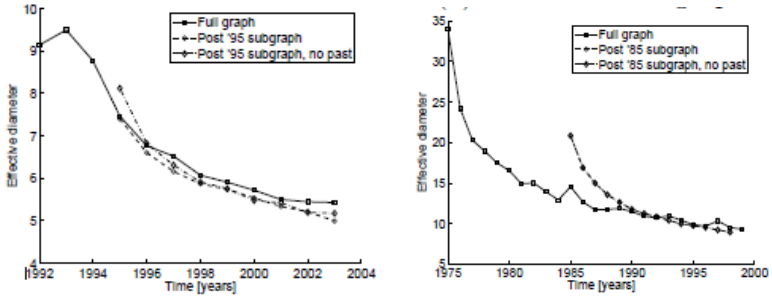
Figure 3: (a) Arxiv (b) Patent

Figure 3(a) shows the DPL plot; the slope is a = 1.68 and corresponds to the exponent in the densification law. Notice that a is significantly higher than 1, indicating a large deviation from linear growth.

Figure 3(b) shows similar pattern as figure (a) with a=1.66.

Shrinking Diameter

The effective diameter used in earlier work: the minimum value *d* such that at least 90% of the connected node pairs are at distance at most *d*. The effective diameter is a more robust quantity than the diameter (defined as the maximum distance over all connected node pairs), since the diameter is prone to the effects of degenerate structures in the graph (e.g. very long chains).



(a) **(b)**

Figure 4: (a) Arxiv Citation Graph (b) Patents

Figure 4 shows the effective diameter over time; one observes a decreasing trend for all the graphs.

Basic Characteristic of Social Network

- Node's Degree – Number of nodes incident to each node
- Network Diameter – Maximum distance between pairs of node
- Effective Diameter – Minimum distance d, such that for at least 90% of reachable node pairs.
- Average Diameter – Other node-to-node distance

2. GRAPH MINING

Graph Mining can be considered a specialization of Data Mining, the objective of the latter being to process data which is difficult for humans to meaningfully interpret, and identify/extract high value knowledge from the data. For example in data mining application analyze the data by the techniques which are in general statistical analysis and/or machine learning techniques using artificial intelligence. Thus graph mining similar to data mining except that it is applied to graph.

3.1 Graph Mining Methods

3.1.1 Apriori-Based Approach

Apriori-based frequent substructure mining algorithms share similar characteristics with Apriori-based frequent item set mining algorithms. The search for frequent graphs starts with graphs of small "size," and proceeds in a bottom-up manner by generating candidates having an extra vertex, edge, or path.

The general framework of Apriori-based methods for frequent substructure mining is outlined in algorithm below. *Sk*is the frequent substructure set of size *k*. At each iteration, the size of newly discovered frequent substructures is increased by one. These new substructures are first generated by joining two similar but slightly different frequent sub graphs that were discovered in the previous call to Apriori Graph. The frequency of the newly formed graphs is then

checked. Those found to be frequent are used to generate larger Candidates in the next round.

However, the candidate generation problem in frequent substructuremining is harder than that in frequent item set mining, because there are many ways to join two substructures.

Algorithm

Input:

D, a graph data set;

min sup, the minimum support threshold

Output:

Sk, the frequent substructure set

*S*1 frequent single-elements in the data set;

Call Apriori Graph(*D*, *min sup*, *S*1);

Procedure Apriori Graph(*D*, *min sup*, *Sk*)

- *Sk*+1 ?;
- for each frequent *gi*2 *Sk*do
- for each frequent *gj*2 *Sk*do
- for each size (*k*+1) graph *g* formed by the merge of *gi*and *gj*do
- if*g* is frequent in *D* and *g* 62 *Sk*+1 then
- insert *g* into *Sk*+1;
- if*sk*+1 6= ? then
- *Apriori Graph*(*D*, *min sup*, *Sk*+1);
- return;

3.1.2 Pattern-Growth Approach

The Apriori-based approach has to use the breadth-first search (BFS) strategy because of its level-wise candidate generation. In contrast, the *pattern-growth approach* is more flexible regarding its search method. It can use breadth-first search as well as depth-first search (DFS), the latter of which consumes less memory.

A graph *g* can be *extended* by adding a new edge *e*. The newly formed graph is denoted by $g\Diamond_x e$. Edge *e* may or may not introduce a new vertex to *g*. If *e* introduces a new vertex, we denote the new graph by $g\Diamond_{xf} e$, otherwise, $g\Diamond_{xb} e$, where *f* or *b* indicates that the extension is in a *forward* or *backward* direction.

A general framework for pattern-growth frequent sub-structure mining is illustrated below. It is simple but not efficient. Same graph can be discovered many times makes it inefficient.

Algorithm

Input:

g, a frequent graph;

D, a graph data set;

min sup, minimum support threshold

Output:

The frequent graph set, *S*

S←Ø

Call Pattern Growth Graph(*g*, *D*, *min_sup*, *S*);

Procedure Pattern Growth Graph(*g*, *D*, *min_sup*, *S*)

- if*g* ∈*S* then return;
- else insert *g* into *S*;
- scan *D* once, find all the edges *e* such that *g* can be extended to $g\Diamond_x e$;
- for each frequent $g\Diamond_x e$do
- *Pattern Growth Graph*($g\Diamond_x e$, *D*, *min sup*, *S*);
- return;

3.2 Application of Graph Pattern Mining

- Mining biochemical structures
- Finding biological conserved sub networks
- Finding functional modules
- Program control flow analysis
- Intrusion network analysis
- Mining communication networks
- Anomaly detection
- Mining XML structures
- Building blocks for graph classification, clustering, compression, comparison, correlation analysis, and indexing

3.3 Graph Generation Models

- **Random Graphs**
  - Gives few components and small diameter

- Does not give high clustering and heavy-tailed degree distributions
- Is the mathematically most well-studied and understood model

- **Watts-Strogatz Model**
  - give few components, small diameter and high clustering
  - does not give heavy-tailed degree distributions
- **Scale-Free Networks**
  - Gives few components, small diameter and heavy-tailed distribution
  - Does not give high clustering
- **Hierarchical Networks**
  - few components, small diameter, high clustering, heavy-tailed

3. CHALLENGES IN CRAWLING

Crawling the Entire Connected Graph

The primary challenge in crawling large graphs is covering the entire connected component. In the case of online social networks, crawling the graph efficiently is important since the graphs are large. Common algorithms for crawling graphs include breadth-first search (BFS) and depth-first search.

Speed of Crawling

As social network are highly dynamic it changes over very quickly. So nodes and edges are keep adding and removing from the graph.

Type of Graph

There are directed and undirected graph. So crawling directed graph, as opposed to undirected graph, presents additional challenges. Many graph can be crawled by only using forward links. But it does not crawl the entire graph instead it explores connected component reachable from set of seed node.

5. CONCLUSIONS

As this paper studies about the basic of graph theory, graph mining and social networking. It has also presented some of the characteristics of social networking. Now it can be learned that social network uses scale-free network model to built graph. It has also presented case study on the analysis of Facebook site and measure the some of the metrics and compared with two sampling techniques.

REFERENCES

1. J. Diesner, T. L. Frantz, and K. M. Carley. Communication networks from the enron email corpus "it's always about the people. enron is no different". Comput. Math. Organ. Theory, 11(3):201–228, 2005.
2. J. Leskovec and E. Horvitz. Planetary-scale views on a large instant messaging network. In WWW ’08:

Proceeding of the 17th international conference on World Wide Web, pages 915–924, New York, NY, USA, 2008. ACM.

3. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta,S. Mukherjea, and A. Joshi. On the structural properties of massive telecom call graphs: findings and implications. In CIKM ’06: Proceedings of the 15th ACM international conference on Information and knowledge management, pages 435–444, New York, NY, USA, 2006. ACM.
4. M. Newman, A.-L. Barabasi, and D. J. Watts, editors. The Structure and Dynamics of Networks. 2006.
5. David F Nettleton, Data mining of social network represented as graph, Dec-2012.
6. R. Albert, H. Jeong, and A.-L. Barabasi. Diameter of the world-wide web. Nature, 401:130–131, September 1999.
7. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: experiments and models. In Proceedings of World Wide Web Conference, 2000
8. S. Milgram. The small-world problem. Psychology Today, 2:60–67, 1967.
9. D. J. Watts, P. S. Dodds, and M. E. J. Newman. Collective dynamics of ’small-world’ networks. Nature, 393:440–442, 1998
10. jureleskovec, john Kleinberg, Christos.Graph Evolution: Densification and Shrinking Diameters.
11. M. Gjoka, M. Sirivianos, A. Markopoulou, and X. Yang. Poking facebook: characterization of osn applications. In Proceedings of the first workshop on online social networks, pages 31{36. ACM, 2008.
12. S A Catanese, Pasquale De Meo, Emilio Ferrara, G Fiumara, A Provetti, Crawling Facebook for Social Network Analysis Purposes, May-2011.
13. R. Albert. Diameter of the World Wide Web. Nature, 401(6749):130, 1999.
14. R. Albert and A. Barab_asi. Statistical mechanics of complex networks. Reviews of modern physics, 74(1):47{97, 2002.
15. J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pages 177{187. ACM, 2005.
16. Wen-jun, S., and Hang-ming, Q. “A Social Network Analysis on Blogspheres” In Proceedings of the 15th IEEE International Conference on Management Science and Engineering, 2008, pp.1769 - 1773, Long Beach, CA, USA.