Analysis of social network & an approach towards evaluation of spreading of epidemics in randomized social Network.

Rashmita Panigrahi[1], Trilochan Rout[2] and Manas Ranjan Mishra[3]

[1] Asst. Prof in CSE.
Computer Science and Engineerinmg .
NM Institute of Engineering and Technology.
Bhubaneswar, Odisha
INDIA

[2] Asst. Prof in CSE.
Computer Science and Engineerinmg .
NM Institute of Engineering and Technology.
Bhubaneswar, Odisha
INDIA

3 Asst. Prof in CSE.
Computer Science and Engineerinmg .
C. V. Raman College of Engineering
Bhubaneswar, Odisha, INDIA

Abstract: Classifying nodes in a network is a task with wide range of applications .it can be particularly useful in epidemics detection .Many resources are invested in the task of epidemics and precisely allow human investigators to work more efficiently. This work creates random and scale- free graphs the simulations with varying relative infectiousness and graph size performed. By using computer simulations it should be possible to model such epidemic Phenomena and to better understand the role played by the different parameters. Such simulations can then be used to study potential measures that can be taken to prevent or at least hinder the epidemic from spreading. Four different type s of network were used represent different structures of connection between individuals: random, scale-free, small-world and small-world with weighted edges. We have varied several free parameters of the graphs and diseases to study their role in disease spreading, for example to findout whether the speed of disease spreading and the number of affected

individuals will change. The size of the network has also been modified to determine how epidemics in small networks relate to epidemics in large ones. Variations in infectiousness of the disease have also been studied to see if there is some minimum value of this parameter under which the epidemic does not spread and to determine what effects it changes have on spreading of the disease. Then the latency period of the disease has been have on spreading of the disease. Then the latency period of the disease has been varied in an attempt to observe differences between diseases with and without a latency Period. Finally simulations for different mean degrees of the network have been performed to determine an effect on characteristics of epidemics.

We have implemented a simple model of epidemics spreading in networks described by arbitrary graphs. We have then performed simulations to study how certain characteristics of the epidemics depend on the structure of the network and several parameters of our model. The four types of networks used in this paper: random networks, scale-free networks according to unweighted and weighted small-world networks.

Keywords: Social Network Analysis (SNA), Social Network Analysis for risk evaluation (SNARE), General leadger 1(GL1), General leadger 2(GL2), Computer Mediator communication (CMC), Internet Relay chat (IRC).

I. INTRODUCTION

Networks have been studied as graphs in mathematics, physics, sociology, engineering and computer science, biology and economics. Each field has its own theory of networks and each field has its own way of aggregating collective behavior. So why is this new? In the past, the networks have been viewed as objects of pure structure whose properties are fixed in time. Both these assumptions are far from truth. Real networks represent populations of individual components that are actually doing something-involved in communication, generating power, sending data, or even making decisions.

Social network analysis (SNA) is a set of research procedures for identifying structures in systems based on the relations among actors. Grounded in graph and system theories, this approach has proven to be a powerful tool for studying networks in physical and social worlds, including on the web [3, 4, and 5] SNA focuses on relations and ties in studying actor's behavior and attitudes. Thus the positions of actors within a network and the strength of ties between them become critically important. Social position can be evaluated by finding the centrality of a node identified through a number of connections among network members. Such measures are used to characterize degrees of influence, prominence and importance of certain members [6]. Tie strength mostly involves closeness of bond.

There is general agreement that strong ties contribute to intensive resource exchange and close communities, whereas weak ties provide integration of relatively separated social groups into larger

social networks [7, 8]. The notion of a social network and the methods of social network analysis have attracted considerable interest from the social and behavioral and computer science community in recent decades. Much of this interest can be attributed to the appealing focus of social network analysis on relationships among social entities, and on the patterns and implications of these relationships. From the view of social network analysis, the presence of regular patterns in relationship, are referred as structure and the quantities that measure structure

The focus on relations, and the patterns of relations, requires a set of methods and analytic concepts that are distinct from the methods of traditional statistics and data analysis.

1.2 Motivation

Social Networking web services are internet applications that help connect friends, business partners or just about any one by allowing users to have profiles, blogs interact with others, join or create communities and much more. In simple models, the rate of epidemic spreading is often represented by the so- called basic reproduction number R0 which is the mean number of secondary cases caused by a single infected individual. In general, R0 varies a lot depending on the geographical location (particularly on people's life style), the season, the density of population and other parameters. In this project, we try to develop a simple epidemic model in graphs which also takes into account the network structure and several other parameters. We believe that by using more four different types of network were used in our project to represent different structures of connections between individuals: random, scale-free, small-world and small-world with weighted edges. We have varied several free parameters of the graphs and diseases to study their role in disease spreading, for example to find out whether the speed of disease spreading and the number of affected individuals will change. The size of the network has also been modified to determine how epidemics in small networks relate to epidemics in large ones. Variations in infectiousness of the disease have also been studied to see if there is some minimum value of this parameter under which the epidemic does not spread and to determine what effects its changes have on spreading of the disease. Then the latency period of the disease has been varied in an attempt to observe differences between diseases with and without a latency period. Finally, simulations for different mean degrees of the network have been performed to determine its effect on characteristics of the epidemic.

2. **Social network analysis:**

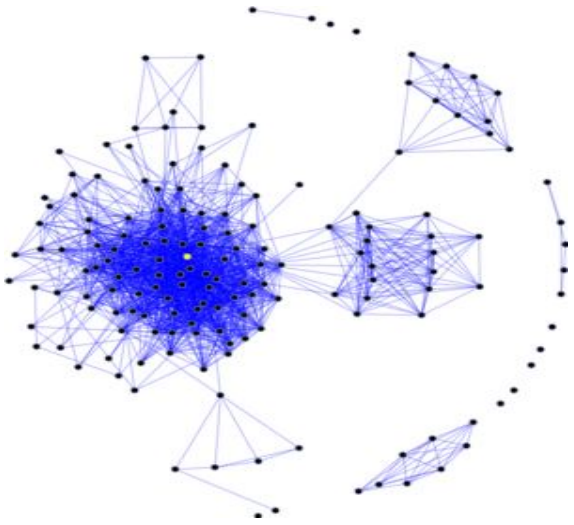

Fig.2 .1

An example of a social network diagram. The node with the highest between ness centrality is marked in yellow. Social network analysis (related to network theory) has emerged as a key technique in modern sociology. It has also gained a significant following in anthropology, biology, communication studies, economics, geography, information science, organizational studies, social psychology, and sociolinguistics, and has become a popular topic of speculation and study.

Social network analysis has now moved from being a suggestive metaphor to an analytic approach to a paradigm, with its own theoretical statements, methods, social network analysis software, and researchers. Analysts reason from whole to part; from structure to relation to individual; from behavior to attitude. They typically either study whole networks (also known as complete networks), all of the ties containing specified relations in a defined population, or personal networks (also known as egocentric networks), the ties that specified people have, such as their "personal communities". In the latter case, the ties are said to go from egos, who are the focal actors who are being analyzed, to their alters. The distinction between whole/complete networks and personal/egocentric networks has depended largely on how analysts were able to gather data. That is, for groups such as companies, schools, or membership societies, the analyst was expected to have complete information about who was in the network, all participants being both potential egos and alters. Personal/egocentric studies were typically conducted when identities of egos were known, but not their alters. These studies rely on the egos to provide information about the identities of alters and there is no expectation that the various egos or sets of alters will be tied to each other.

3. PROPOSED METHOD:

Given:

A graph G = (V, E), where entities (persons, accounts, blogs, etc.) are represented as vertices, or nodes, in the graph, and interactions (phone calls, account trans- actions, hyperlinks) between them are represented as edges.

• Binary class (state) labels X = {x1 , x2 } defined on V .

• A set of flags for each node vi, V, based on node attributes (geographic location, name, etc.)

Output: A mapping V → X from nodes to class labels.

The labels X are binary categorical variables derived from the context (normal or irregular, conservative or liberal, etc.). We also note that while nodes and links can be related to social entities such as persons and relations or actions, the proposed methods can be applied to any sort of entities, such as accounts or web-pages.

The basic premise is to use neighboring labels to classify a given node. This premise has proven effective for many graph labeling tasks. However, we also take into account domain knowledge, by assigning an initial risk scores to nodes prior to evaluating neighborhood associations between them. To measure risk by association, we then use belief propagation for passing risk to connected nodes. A detailed tutorial of belief propagation may be found in work by Yedidia.

Let us summarize the procedure. In a network for a given task, the true label for each node vi is

unknown. We are, however, given some local observations about the node, which we use as a local estimation of its risk, or node potential φi (xc) of vi for class xc (the procedure for determining this will be described shortly). Information about this node is inferred from the surrounding nodes. This is obtained through iterative message passing to and from vi to each neighbor vj , where a message from vi to vj with its own assessment of vj 's believed class is denoted by mij . At the end of the procedure, the belief of a node vi belonging to in class xc is determined. The belief is an estimated probability, which can be threshold into the classes (e.g. a bi (xc) > .5 implies vi belongs to class xc), or used relatively to compare risk scores between nodes (e.g. bi (xc) > bj (xc) implies vi is more likely to belong to xc than vj).

In more detail, messages are obtained the following way. Each edge eij has associated messages mij (xc) and mji (xc) for each possible class. mij (xc) is a message that vi sends to vj about vj believed likelihood of belonging to xc . Iter-atively, messages are updated using the sum-product algo-rithm. Each outgoing message from a node to a neighboris updated according to incoming messages from the node's other neighbors. Formally, the message-update equation is as follows:

$$m_{ij}(x_c) \leftarrow \sum_{x_d \in X} \emptyset_i x_d \varphi_{ij}(x_d, x_c) \pi_{l_k \in N_{i/j}} m_{l_i}(x_d)$$

where N (vi) is the set of neighboring nodes to vi . ψij (xc , xd) is the edge potential of an edge between two nodes i, j of classes xc and xd . ψij (xc , xd) is generally large if edges be- tween xc and xd occur often, and small if not. Order of message-passing does not matter, provided all messages are passed in each iteration. We also normalize mij (xc) to avoid numerical underflow, as discussed in [8], so each edge's mes-sage vector sums to one:c mij (xc) = 1.

Case Studies:

We developed an algorithm to help detect risks in accounting data, so we will primarily evaluate it on its ability to find misstated accounts in a company's general ledger. How-ever, since our G/L data is proprietary, and because we believe it is more generally useful, we also evaluate its performance for graph labeling using public data from social media and political campaigns.

Detecting misstated general ledger ac-counts

The general ledger (G/L) of a company is an accounting record that summarizes its financial activity with double- entry bookkeeping. Within every G/L is a set of accounts which can be thought of as variables representing the allocation of monetary resources. Business events, such as the purchase of machinery, would result in a transaction that reduces the value of the cash account but increases the value in the fixed asset account by an equivalent amount.

The G/L is used to prepare the financial statements by aggregating the balances of the accounts and thus auditors are extremely interested in finding misstatements in this data. Manipulation of records can be found by experts on both the G/L and financial statement level. There are many different fraud schemes for which experts have identified "red flags" that indicate suspicious behavior based on domain knowledge. For example, one fraud

scheme is known as channel stung. In order to meet earnings expectation, fictitious sales are recorded to increase the revenue for the current quarter. These sales are typically not complete and are recorded solely to meet the earnings target. The company overloads their distribution channels to make it appear as if additional sales have been completed. This helps the company appear to meet its target. Such channel stung is usually followed by an increase in the number of returns at the beginning of the next quarter. In the general ledger, one could record the return of a sale by debiting revenue and crediting accounts receivable; thus to look for channel stung one might create a threshold test or red flag that highlights an account when there are an excessive number of these transactions.

To analyze general ledger data with SNARE we first need to create a network with nodes, edges, and initial risks. For our application, we construct the network as follows:

• Each account in the general ledger becomes a node in the network.

• For every pair of accounts (X, Y) in the general ledger, they are connected with an edge if there are transactions where the sum of the amounts debiting X and crediting Y exceeds a minimum threshold.

• The initial risks on the nodes is determined by per-forming a preliminary scan over the data to detect red flags as determined by domain experts. The red flags mine the initial risk as defined.

GL1:

In the first set of G/L data there were a total of 1, 380 accounts, 3, 820 edges, and 11, 532 red flags (nearly every node had at least one flag). From prior domain knowledge, 26 accounts were identified as being misstated. We applied SNARE to this network and the message-passing process converged after 6 iterations. Our initial node potentials were φi (Risky) = 0.1 and φi (N otRisky) = 0.9 for a node i with no flags, and additional flags changed node potential accord-ing to Equation 3, so key information is in the nodes' number of flags relative to each other.

GL2:

The second set of G/L data contained 1, 678 nodes, 18, 720 edges, and 11, 401 red flags. Unfortunately, with this data set we had only coarse label information available that identified general groups of misstated accounts. For our experiments we treated all accounts in an identified group as being misstated, resulting in a total of 337 positive labels.

4. Implementation:

In our implementation of the above described model we represented the graphs by their adjacency matrices. Here we present the scripts for MATLAB that we used to generate them

Random graphs:

```
1.  function graph = create_graph_rnd(N,d);
2.  graph = zeros(N,N);
3.  for i = 1:(d*N/2)
4.  j = floor(N*rand)+1;
5.  k = floor(N*rand)+1;
6.  while (j==k)||(graph(j,k)==1)
7.  j = floor(N*rand)+1;
8.  k = floor(N*rand)+1;
9.  end;
```

```
10.  graph(j,k)=1;
11.  graph(k,j)=1;
12.  end;
13.  graph = sparse(graph);
```

We have used the short script shown above to generate random graphs with a given number of vertices N and a given mean degree d. The N d edges are placed between randomly chosen pairs (j, k) of vertices. If a pair is chosen that is already connected or if j = k then the selection step is repeated (lines 7–10). Finally, the resulting adjacency matrix is stored as a sparse matrix to save memory (line 13).

Small-world graphs:

The method presented by Watts and Strogatz [1998] is used to create a graph with a given number of vertices N and a mean degree d. To gain time and because the matrix is symmetric (i.e. if i is connected to j, then j is also connected to i), the building process is only done on one half of the matrix, the upper right side.

```
1.  function graph = create_graph_sw(N,degree,p)
2.  graph = sparse(zeros(N,N));
3.  N_initial_edge=floor(floor(degree+.5)/2);
4.  for i = 1:N
5.  for link = 1:N_initial_edge
6.  graph(i,mod(i+link-1,N)+1)=1;
7.  end;
8.  end;
9.  k=floor(degree+.5)/2;
10. if(floor(k)<k)
11. for i =1:2:N
12. graph(i,mod(i+floor(k),N)+1)=1;
13. end;
    end;
14. for i=1:N
15. for j=i+1:N
16. if(graph(i,j)==1 && rand()<p)
17. graph(i,j)=0;
18. a=floor(rand()*N)+1;
19. while(graph(a,i)==1 || graph(i,a)==1 || i==a)
20. a=floor(rand()*N)+1;
21. end;
22. if(i<a)
23. graph(i,a)=1;
24. else
25. graph(a,i)=1;
26. end;
27. end;
28. end;
29. end;
```

In this script we first calculate the number of edges corresponding to a mean degree d′ given by the natural number which is closest to the given value d. The next step is to create a regular network with vertices arranged in a circle where every vertex is connected to its d′ closest neighbours (lines 3–15). More precisely, each vertex is linked with the d′ /2 following neighbours in the network. In this way every vertex will in the end have 2d′ /2 connections (lines 3–9). If the mean degree d′ is odd, every second vertex is then connected to its d′ /2 th next neighbor (lines 10–15).

5. Epidemic simulation

The following function epidemic step takes as its parameters the state vector
old states, the adjacency matrix graph, a vector disease containing the infec-

tiousness values α1 , α2 , . . . , αL of the different stages of the disease and the value of relative infectiousness κ and returns the new state vector new states after one epidemic step.

```
1. function new_states =
   epidemic_step(old_states, graph, disease, k);
2. infectiousness = zeros(length(old_states),1);
3. for individual = 1:length(old_states)
4. if (old_states(individual) > 0)
5. infectiousness(individual) =
   disease(old_states(individual));
6. end;
7. end;
8. prob = (graph*infectiousness)*k;
9. for individual = 1:length(old_states)
10. if (old_states(individual) > 0)
11. if (old_states(individual) == length(disease))
12. new_states(individual) = -1;
13. else
14. new_states(individual) =
   old_states(individual) + 1;
15. end;
16. else
17. if (old_states(individual) == 0)
18. if (rand<prob(individual))
19. new_states(individual) = 1;
20. else
21. new_states(individual) = 0;
22. end;
23. else
24. new_states(individual) = -1;
25. end;
26. end;
27. end;
```

First an N-dimensional vector infectiousness is created whose elements are either 0 for non-infected individuals or equal to the value αs(j) of the infectiousness corresponding to the disease stage s(j) in which the particular individual is. Then the probabilities of infection for each individual are calculated according to the relation (1) and stored in a vector prob (line 11).

Afterwards, infected individuals are evolved deterministically either to the next stage of the disease (line 17) or – if they have already reached the last stage – to the resistant state (line 15). Susceptible individuals are infected with probabilities given by the vector prob (lines 21–25) and resistant individuals remain resistant (lines 26–27).

6. Conclusion:

Availability of realistic and accurate epidemic models is undeniably important for our better understanding of epidemics and the way they spread. The possibility of inexpensive experimenting with parameters of the model is essential for development of efficient defense strategies against contagious diseases. Therefore we believe that epidemic modeling is a perspective field which will further evolve but this will surely require extensive collaboration with social sciences which could provide better understanding of the large scale structure of human social network.

Another interesting possibility for further work on this topic could in our opinion be simulations in lattice-based small world networks. A big advantage of small world graphs is that they have a well defined regular underlying structure. This means that in principle one does not need to store the full adjacency matrix but only information about reconnected edges. As a result, epidemics could be simulated in large networks using a cellular automaton model with connections between nearest neighbours in the lattice and few "long distance" edges which could be either fixed or simply randomly generated at each step. Such model would include both local spreading of the epidemic as well as transfer over longer distance mediated by traffic.

References:

[1] L. A. N. Amaral, A. Scala, M. Barth´l´my, and H. E. Stanley. Classes of small worldee Networks proc. Natl. Acad. Sci USA 97:11149 2000.

[2] Albert-L´szl´ Barab´si and R´ka Albert. Emergence of scaling in random networks.aoae Science, 286:509, 1999

[3] Francesc Comellas and Michael Sampels. Deterministic small-world networks. Phys ica A, 309:231, 2002.

[4] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' net works. Nature, 393:440, 1998.

[5] Duncan J. Watts. Six Degrees: The Science of a Connected Age. W. W. Norton & Company.

[6] S.D. Berkowitz. An Introduction to Structural Analysis: The Network Approach to Social Research. Toronto: Butterworth

[7] S. Bay, K. Kumaraswamy, M. G. Anderle, R. Kumar and D. M. Steier. Large scale detection of irregularities in accounting data. In ICDM '06 Proceedings of the Sixth International Conference on Data Mining, pages 75–86, Washington, DC, USA, 2006. IEEE Computer Society.

[8] R. Behrman and K. Carley. Modeling the structure and effectiveness of intelligence organizations: Dynamic information flow simulation. In Proceedings of the 8th International Command and Control .Research and Technology Symposium., 2003.

[9] T. Bell and J. Carcello. A decision aid of assessing the likelihood of fraudulent financial reporting. Auditing: A journal of practice and theory, 19:169–184, 2000.

[10] R. Bolton and D. Hand. Statistical fraud detection: A review, 2002.

[11] T. Cohn. Scaling Conditional Random Fields for Natural Language Processing. PhD thesis, University of Melbourne, 2007.