Spread of Influence Across Disconnected Communities

Arushi Khattri

A dissertation submitted for the partial fulfilment of BS-MS dual degree in Science

Indian Institute of Science Education and Research Mohali
April 2014

Certificate of Examination

This is to certify that the dissertation titled **Spread of influence across disconnected communities** submitted by **Arushi Khattri** (Reg. No. MS09024) for the partial fulfilment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. C. S. Aribam Dr. Lingaraj Sahu Dr. V. R. Srinivasan

(Supervisor)

Dated: April 25, 2014

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Dr.Varadharaj R. Srinivasan at the Indian Institute of Science Education and Research Mohali.

This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgment of collaborative research and discussions.

Arushi Khattri
(Candidate)

Dated: April 25, 2014

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Varadharaj R. Srinivasan
(Supervisor)

Acknowledgment

I would like to thank my supervisor Dr.Varadharaj R. Srinivasan for his guidance and support throughout my thesis work. I am truly grateful towards him for his valuable time and an opportunity to learn with him. I want to address a very special thanks to Dr. Sudarshan Iyengar (IIT Ropar) for acting as my co-supervisor and mentoring me throughout my thesis. Without their guidance and support none of the work presented in this thesis would have been possible. I am also thankful to the Department of Mathematical Sciences, IISER Mohali for giving me an opportunity to write this thesis.

Finally, I would like to express my deepest gratitude to my family and friends for being there by my side and giving me all the support needed.

List of Figures

Abstract

In social networks, communities are defined by a group of highly intra-connected and sparsely inter-connected nodes in a network. Detecting Communities, dynamics of its formation and and understanding its relevance in social network analysis has been of great interest across several disciplines such as sociology, mathematics, computer science, physics and Epidemiology.

The current thesis is an attempt to understand distances between communities and use the results thus obtained in analysing the reservation system that is prevalent in India from the past 7 decades.

We model the the problem in network theory terms and study the distinct community formation that takes place in the Indian society based on caste-based homophily. Reservation system has been in practice since the independence and instead of social upliftment this system is believed to have caused a social disparity amongst the socially forward and the socially backward classes. We model the reservation system and show that it has played an important role in reducing the distance between the uplifted and the downtrodden, hence drawing a bridge between the disconnected classes. We define the term spread of influence in terms of the average shortest path and study the changes in the average shortest path when bridges are added between the clusters. We present our results empirically and make an attempt to give a theoretical explanation for it.

Contents

Chapter 1

A network is a set of nodes or vertices joined together in pairs by links or edges. Many systems could be represented in the form of a network, for example technological networks like the World Wide Web, the Internet and biological systems like the metabolic network, neural network, gene regulatory network and social networks like the friendship network. Network theory is a "language for talking about networks that is precise enough to describe not only what a network is but what kinds of different networks there are in the world"[J.03]. Over the past six decades network theory has been applied in many different domains starting from sociology to biology to physics. In 1937, a psychologist Jacob Moreno, a German psychology researcher, first developed the concept of social networks. He developed a network model to analyze human social groups by studying a group and finding whether the psychological state of an individual is dependent on the relationships between the group. Moreno developed the "Sociogram", a diagrammatic representation of the relationships between individuals in a social group. Then after a huge gap of twenty years, a paper by Catwright and Harary(1956) claimed that the Sociograms developed by Moreno could be analyzed using graph theory. This marked the very beginning of network analysis. Then in the late 1960s, Stanley Milgram, a Harvard sociology professor, performed an experiment to investigate the unresolved hypothesis of "the small world phenomenon" which proved to be major step in network theory. This theory claimed that the world is in a sense small, as when viewed as a network of social acquaintances, one could reach any other individual in a few steps through the network of friends. This led to the phrase "six degree of separation"[Mil67]. Then the question of "What would it take for any world to be small?" came up and people felt the need to model this using some mathematical tools. Then Erdős and Rényi introduced the theory of random graphs using which one could explain a lot of facts of the network theory by randomly constructing networks.

In this chapter we shall give an overview of network analysis by discussing some basic concepts about network theory followed by discussing the community structures in networks and the algorithms to detect them. Later we shall discuss the types of networks studied and in the end give an introduction to the social networks.

1.1 Preliminaries

Suppose we have a network $N = (V, E)$ where V is the set of vertices or nodes and E is the set of links or edges. The network N is *directed* if the edges in the network have a direction i.e. they start from one node and terminate at another and N is *undirected* if there is an edge between two nodes without a source and destination. Let $u \in V$, the *degree* k_u of the node u is defined as the number of links the node u has to other nodes. For a directed network, the number of incoming links to a node is called the *indegree* of the node and the number of outgoing nodes is called the *outdegree*. A *path* between two nodes $u \in V$ and $v \in V$ in a network is defined as the sequence of edges from u to v. The *eccentricity* of a node u is the maximal shortest path from u to all the other nodes. The *diameter* of a network N is the maximum eccentricity across all nodes and the *radius* is the minimum eccentricity across all nodes.
A network N is represented in form of a matrix, A and its element a_{ij} is defined as:

$$a_{ij} = \begin{cases} 1 & \text{if there is an edge from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

This matrix is called an adjacency matrix. This matrix fully specifies the topology of the network.[FRP04] We can calculate the degree of a node using $k_u = \sum_v A_{uv}$.

1.1.1 Degree Distribution

The degree distribution is the probability distribution of the degree of the nodes over the whole network i.e. if $D(k)$ is the degree distribution of the network N it gives the fraction of nodes of N which have degree k.
Let $x : V \longrightarrow \{1, 2, \ldots, (n-1)\}$ be the degree of a node.
Let $P(x = d) : \{1, 2, \ldots, (n-1)\} \longrightarrow \mathbb{Q}$ be the probability that a node has degree d.
Define:

$$P(x = d) = \frac{\text{Number of nodes with degree } d}{\text{Total number of nodes}}$$

such that $\sum_{d=1}^{(n-1)} P(x = d) = 1$.
With this we can find the probability density function, hence the distribution $D(k)$.

1.1.2 Connected Components

A connected component of an undirected graph, N is an induced subgraph formed by the equivalence class with an equivalence relation defined on the set of nodes. For two nodes $v_1, v_2 \in V$ equivalence relation is defined as $v_1 \sim v_2$ i.e. v_1 is reachable from v_2 if there is a path from v_1 to v_2. This is an equivalence relation as:

- *Reflexive* There is a trivial path from a vertex to itself
- *Symmetric* If there is a path from v_1 to v_2 then the same sequence of edges could be taken as a path from v_2 to v_1 (Since the graph is undirected).
- *Transitive* If there is a path from v_1 to v_2 and from v_2 to v_3 then both the paths can be joined to get a path from v_1 to v_3

For a directed graph, N we can define a strongly connected component(SCC) and a weakly connected component(WCC). A *strongly connected component* is an equivalence class with equivalence relation $v_1 \sim v_2$ if and only if there exists a path $v_1 \to v_2$ and a path $v_2 \to v_1$ in N .A *weakly connected component* is the maximal subgraph of a directed graph such that for a pair of vertices v_1 and v_2 there is a directed path from v_1 to v_2 and an undirected path from v_2 to v_1

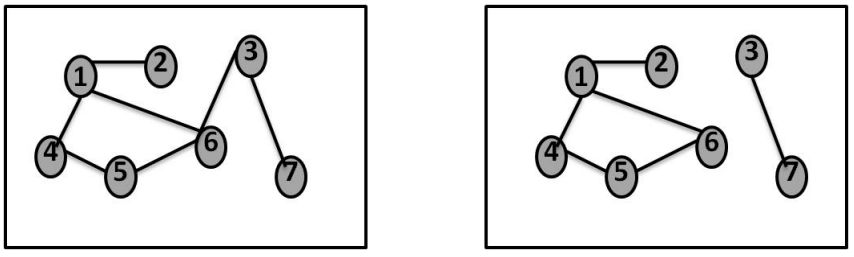

Figure 1.1: A graph that is connected and a graph that contains connected components

1.1.3 Communities

A community is a subset of vertices which are densely intra-connected and sparsely inter-connected with other subsets.A *densely connected* subgraph is defined as the

graph which has more than expected number of edges and similarly a *sparsely connected* graph has less than expected number of edges.[M.E04] In other words, if $H \subset N$ and $u \in V$. Then we can write $k_u(H)$ i.e. the degree of u in the subgraph H as[FRP04]

$$k_u(H) = k_u^{in}(H) + k_u^{out}(H)$$

where $k_u^{in}(H) = \sum_{v \in H} a_{uv}$ and $k_u^{out}(H) = \sum_{v \notin H} a_{uv}$. Then we can say that H is a community if

$$k_u^{in}(H) > k_u^{out}(H) \forall u \in V$$

Communities are observed in may real-world network like the group of web sites related to a particular topic is a community in the World Wide Web network and the biological units performing similar functions in any biological system.

In network analysis it is important to study communities in a network as the net-

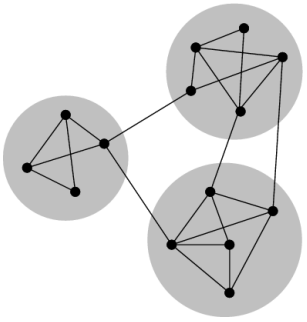

Figure 1.2: Network Representation of a community[New06]

work might have some properties at a community level which are different from the properties of the network globally, position of a node in the community and at the border might affect its role in the network.

The process of community detection can be seen as a mapping of a network into a tree, commonly known as dendogram amongst sociologist as shown in the figure 1.3. [FRP04]

Community Detection Methods

In this section we shall discuss the methods to detect communities in a given network. We shall start by discussing some traditional approaches and then move on to some recent work which is again divided into global approaches and local approaches.

Suppose we have a network N which can be divided into communities such that every

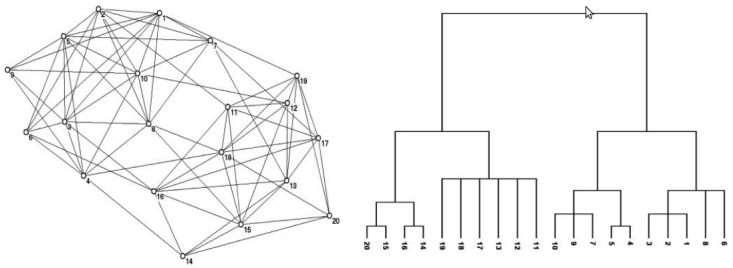

Figure 1.3: A simple network and the corresponding dendogram[FRP04]

node is in one of the communities.

1. **Traditional Approach**

   Traditional approaches partioned the vertex set into communities while minimizing the number of inter-community edges. Spectral Bisection and Kernighan-Lin are the two main algorithms which followed this approach. These algorithms give the best possible division of vertices although it the user has to specify the size of the community and the number of communities required beforehand. Now, we shall discuss both the algorithms in detail.

   - *Spectral Bisection*:
     Let $G = (V, E)$ be an undirected graph on $|V| = n$ vertices and A be an $n \times n$ adjacency matrix of G.Let D be a diagonal matrix where

     $$d_{ij} = \begin{cases} \text{degree of node } i & i = j \\ 0 & \text{otherwise} \end{cases}$$

     The *Laplacian matrix* of G is an $n \times n$ symmetric matrix defined as

     $$L = D - A$$

     Lemma 1.1. *The Laplacian matrix is positive semi-definite.*

     Proof Let $G_{1,2}$ be a graph on two vertices with one edge.Then

     $$L_{G_{1,2}} = D_{G_{1,2}} - A_{G_{1,2}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{1.2}$$

Let x be a vector. Then

$$x^T L_{G_{1,2}} x = (x(1) - x(2))^2 \tag{1.3}$$

Now, for a graph G=(V,E) on n vertices we define

$$L_G = \sum_{(u,v)\in E} L_{G_{u,v}} \tag{1.4}$$

It follows that for all $x \in \mathbb{R}^V$

$$x^T L_{G_{1,2}} x = \sum_{(u,v)\in E} (x(u) - x(v))^2 \tag{1.5}$$

For eigenvector w and eigenvalue λ, we have

$$w^T L_G w = \lambda w^T w \geq 0 \tag{1.6}$$

So, every eigenvalue of L_G is non-negative hence L_G is positive semi-definite.

From equation (1.3) it follows that if all entries of a vector x are same then $x^T L_G x = 0$ which implies that $L_G x = 0$ so such vectors are eigenvectors for eigenvalue 0.

Lemma 1.2. *Let $G = (V, E)$ be a graph with $\lambda_1 = 0 \leq \lambda_2 \leq \cdots \leq \lambda_n$, the eigenvalues of the Laplacian of G. Then $\lambda_2 > 0$ if and only if G is connected.*

Now,let the eigenvalues of L be ordered as $\lambda_1 = 0 \leq \lambda_2 \leq \cdots \leq \lambda_n$. All rows and column of L sum to 0, so the eigenvector corresponding to the eigenvalue λ_1 is $\mathbf{1}_n = (1, 1, \ldots, 1)$. The number of connected components in the graph is given by the multiplicity of λ_1.
If the graph G separates into perfect communities i.e. there are no edges between the communities then the matrix L will be a block diagonal. Each diagonal block will form a Laplacian of a particular component and will have a corresponding eigenvector $v^{(k)}$ with eigenvalue zero, where

$$v^{(k)} = \begin{cases} 1 & i \in G_k \\ 0 & \text{otherwise} \end{cases}$$

Thus, it will have g degenerate eigenvectors with eigenvalue zero.
If the graph does not divide into perfect communities then L will not be in the form of a block diagonal but it will still have the eigenvector $\mathbf{1}_n$ with eigenvalue zero and g-1 eigenvectors with eigenvalue slightly greater than zero (all eigenvalue of a graph Laplacian are non-negative). So, by taking the linear combination of eigenvectors with eigenvalue slightly greater than zero one can approximately find the block diagonals.
If we want to divide the graph in just two communities then we look at the eigenvector,$v^{(2)}$ corresponding to the second lowest eigenvalue(λ_2) and partition the graph as all eigenvectors corresponding to non-degenerate eigenvalues of a real symmetric matrix are orthogonal and all eigenvectors other than that corresponding to eigenvalue zero will have both positive and negative elements.[M.E04]
Let G_1 and G_2 be the two partitions of the graph G. The algorithm to divide the graph G into G_1 and G_2 is as follows:[APL90]

- *Step 1:* Compute the eigenvector $v^{(2)}$ and the median of its components v_m.
- *Step 2:* Partition the vertices of G as for a node $i \in V$, if $v_i^{(2)} > v_m$ put i in G_1 otherwise put it in G_2 .

This method gives a partition of the network. It can be intuitively observed by looking at a vibrating string. When a string is plucked it vibrates at a certain frequency . A sring of length l produces a standing wave at a wave length $\frac{2l}{n}$ for $n \in \mathbb{N}$. We can view the points of the vibrating string as vertices and the string between the two points as an edge, looking at this we can find the Laplacian matrix of this chain graph and the second smallest eigenvalue corresponds to standing wave equal to one whole wavelength. This results in a wave with half points above the equilibrium point and half points below it.The value of the entry for each vertex in the corresponding eigenvector determines the "height" of the vertex above or below the line. Using the sign of this "height as the tool of determining which subgraph to place the vertices in, it is apparent that about half of the vertices would go into each subgraph.[APL90]
The eigenvector $v^{(2)}$ is known as the Fiedler vector and the corresponding eigenvalue λ_2 is known as the algebraic connectivity of the graph G. Smaller

value of λ_2 corresponds to a better partition. Fiedler showed that the two subgraphs obtained are connected.

Theorem 1. *Let G be a connected graph and let $v_{(2)}$ be the eigenvector corresponding to eigenvalue λ_2. For real number $r \geq 0$ define $V_1(r) = \left\{ i \in V : v_i^{(2)} \geq -r \right\}$. Then the subgraph induced by $V_1(r)$ is connected. Similarly for $r \leq 0$, let $V_2(r) = \left\{ i \in V : v_i^{(2)} \leq |r| \right\}$. Then the subgraph induced by $V_2(r)$ is also connected.[M.F73]*

Corollary: *For a graph $G = (V, E)$ if $v_i^{(2)} \neq 0 \forall i \in V$, then the subgraphs induced by both $V_1 = \left\{ i \in V : v_i^{(2)} > 0 \right\}$ and $V_2 = \left\{ i \in V : v_i^{(2)} < 0 \right\}$ are connected.*[M.F73]

In worst-case this algorithm takes $O(n^3)$ time.[M.E04]

- *Kernighan-lin Algorithm*:

  This is a greedy optimization heuristic method which assigns the divisions, a cost function T and optimize it over all possible divisions. In this algorithm we have to specify the size of the communities in which we want to divide the network and give an initial configuration of the community, which could be randomly chosen.

  Let $G = (V, E)$ be a graph, the algorithm find a partition of G into G_1 and G_2. Let C be a weighted connectivity matrix describing the edges of G.

  Let us define for each $a \in G_1$, an *External cost* E_a as

$$E_a = \sum_{v \in G_2} c_{av}$$

  and an *Internal cost* I_a as

$$I_a = \sum_{x \in G_1} c_{ax}$$

  Similarly we define E_b and I_b for $b \in G_2$. Let $D_z = E_z - I_z$ for all $z \in G$ be the difference between the external and internal cost.

  Now, we want to find a partition of G such that the gain in the cost function is maximized. Let h be the total cost due to connections between G_1 and G_2 that do not involve a or b. Then

$$T = h + E_a + E_b - c_{ab}$$

Exchange a and b the new cost T' is

$$T\prime = h + I_a + I_b - c_{ab}$$

so, we have

$$\begin{aligned} Gain &= T - T' \\ &= D_a + D_b - 2c_{ab} \end{aligned}$$

The algorithm finds an optimal series of swaps between nodes in G_1 and G_2 such the *Gain* is maximized.[KL70] The worst-case time for this algorithm is $O(n^2)$ so it is faster than spectral bisection. [M.E04]

2. **Recent Approach**

- *Girvan and Newman Algorithm*: This algorithm detects communities in a network by progressively removing edges from the original network. The measure taken to remove the edges is the edge betweenness value.

$$\text{E.B of an edge } e \in E = \sum_{(i,j)\in V} |d_e(i,j)|$$

where $|d_e(i,j)|$ is the number of shortest path between i and j that run along the edge e. The algorithm is as follows:

(a) Calculate the edge betweenness value of all the edges as
   - For each node i, perform BFS on the graph.
   - Determine the shortest paths from i to all the other nodes and based on that calculate the betweenness value.

(b) Remove the edge with the highest edge betweenness value(say $e = (i,j)$). If two edges have the same value then remove one randomly or remove both of them.

(c) Now recalculate all the edge betweenness value and again remove the edge according to the highest value. Repeat this process as long as there are edges in the graph

In this method the user need not know the size or any property of the community in the network. To know which partition of the network is best we shall optimize

modularity as discussed in the next section. This algorithm takes $O(m^2n)$ time as $O(mn)$ time to calculate the edge betweenness value and recalculates it m times.

Modularity

Modularity is a measure to quantify the quality of a community. Modularity can be positive or negative, higher the modularity value better is the community structure. Modularity is defined as the number of edges falling within groups minus the expected number of edges in an equivalent network where edges are placed at random.
Let $N = (V, E)$ be a network with n vertices and m edges and two groups G_1 and G_2. Define a vector s with n elements as

$$s_i = \begin{cases} 1 & \text{if vertex } i \in G_1 \\ -1 & \text{if vertex } i \in G_2 \end{cases}$$

Let A be the adjacency matrix. Let k_i be the degree of vertex i.

Lemma 1.3. *Let e be the number of edges between i and j if edges are placed at random. Then*

$$E[e] = \frac{k_i k_j}{2m} \tag{1.7}$$

where $m = \frac{1}{2}\sum_i k_i$

Proof Let $e \in E$ be an edge of the network N

$$\begin{aligned} P(\text{e is an edge of node i}) &= \frac{k_i}{2m} \\ P(\text{e is an edge of node j}) &= \frac{k_j}{2m} \\ P(\text{e is an edge of node i and j}) &= \frac{k_i k_j}{(2m)^2} \\ E[e] &= \sum_{a=1}^{2m} \frac{k_i k_j}{(2m)^2} \\ E[e] &= \frac{k_i k_j}{2m} \end{aligned}$$

Let Q be the modularity of the network N and Q_{ij} be the modularity for an edge $(i, j) \in E$

$$Q_{ij} = A_{ij} - \frac{k_i k_j}{2m} \qquad for\, all\, i\, and\, j \tag{1.8}$$

Observe

$$\frac{1}{2}(s_i s_j + 1) = \begin{cases} 1 & \text{if i and j belong to the same group} \\ 0 & \text{if otherwise} \end{cases}$$

Then,

$$Q = \frac{1}{4m} \sum_{i,j} \left[\left(A_{ij} - \frac{k_i k_j}{2m} \right) (s_i s_j + 1) \right] \tag{1.9}$$

Since,

$$2m = \textstyle\sum_i k_i = \sum_{i,j} A_{ij}$$

We have

$$Q = \frac{1}{4m} \sum_{i,j} \left[\left(A_{ij} - \frac{k_i k_j}{2m} \right) (s_i s_j) \right] \tag{1.10}$$

In matrix notation

$$Q = \frac{1}{4m} s^T B s \tag{1.11}$$

where B is a real symmetric matrix known as the modularity matrix with elements

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$$

The rows and column of the matrix B sum up to zero so it has $(1, 1, 1, \ldots, 1)$ as an eigenvector with eigenvalue zero. Now, write s as a linear combination of the normalized eigenvectors u_i of B such that $s = \sum_{i=1}^{n} a_i u_i$ with $a_i = u_i^T.s$. Then we have

$$\begin{aligned} Q &= \frac{1}{4m} \sum_i a_i u_i^T B \sum_j a_j u_j \\ &= \frac{1}{4m} \sum_{i=1}^{n} (u_i^T.s)^2 \beta_i \end{aligned}$$

where β_i is the eigenvalue corresponding to the eigenvector u_i.
Let $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$ be the eigenvalues of the matrix B arranged in decreasing order. We want to maximize the modularity by choosing an appropriate division of the network or equivalently by choosing the index vector s such the most weight is put with the eigenvector having the maximum eigenvalue. If there were no other constraints,

we could have chosen s to be parallel to u_1 as β_1 is the maximum eigenvalue. But as we defined, s can have values $+1$ or -1. So, keeping the constraints in mind, we try to choose s such that it is as parallel to u_1 as possible. To do so, we have to maximize the dot product $u_1^T.s$. It can be easily seen that the maximum is achieved when we put

$$s_i = \begin{cases} 1 & \text{if corresponding element of } u_1 \text{ is positive} \\ -1 & \text{otherwise} \end{cases}$$

In other words, nodes with corresponding elements positive in u_1 go in one group and others in another. This method can be used to find the division of a network without knowing the size of the community beforehand.[New06]

1.1.4 Types of Networks

In this section we shall discuss the types of networks studied so far.

Random Network

A *Random network* is a collection of nodes with edges placed between them at random. [R84] Let $G_{n,m}$ be a set of graphs with n labelled nodes and m edges. A graph $g \in G_{n,m}$ is formed by choosing m out of the $N = \binom{n}{2}$ edges. Therefore there are $\binom{N}{m}$ elements in $G_{n,m}$. So, a random graph $g_{n,m}$ can be defined as an element of $G_{n,m}$ which is chosen with a probability $\frac{1}{\binom{N}{m}}$. This is known as the $G(n,m)$ model.[Bol98]
Another model is defined on a set of graphs, $G_{n,p}$ where n is the number of nodes in the graph and an edge is placed between two nodes with a probability $0 \leq p \leq 1$. Probability for graph $h(n,m)$ i.e. a graph on n nodes with m edges is $p^m(1-p)^{N-m}$) where $N = \binom{n}{2}$.
The spaces $G_{n,m}$ and $G_{n,p}$are closely related as in if we condition the $edges(G_{n,p}) = m$ then we obtain $G_{n,m}$.[Bol98] So, from now on we shall assume the random graph to be the $G_{n,p}$ model. Now, let us calculate the degree distribution for this network.
First, we shall calculate the expected number of edges in a random graph($\Upsilon(V,E) = g_{n,p}$).
Let $e : V \times V \longrightarrow \{0,1\}$ be an indicator random variable.
defined as

$$e(i,j) = \begin{cases} 1 & \text{if there is an edge between i and j} \\ 0 & \text{otherwise} \end{cases}$$

we know by definition of random graphs that $P[e = 1] = p$.
So,the expected number of edges in $\Upsilon(V, E)$ is

$$\begin{aligned} E[e] &= \sum_{i=1}^{n} \sum_{j=i}^{n} ep \\ &= p\binom{n}{2} \\ &= pN \end{aligned}$$

Define k_i to be the degree of a node i.Let $\langle k \rangle$ be the *mean degree* of the graph i.e. the mean number of edges attached to a node. Then

$$\langle k \rangle = \frac{2E[e]}{n}$$

where the factor 2 comes as each edge is attached to two nodes. Thus,

$$\langle k \rangle = \frac{2p\binom{n}{2}}{n} = \frac{2pn(n-1)}{2n} = pn(n-1) \simeq pn$$

Now, to find the probability that a node has degree k.
Define $x : V \longrightarrow \{1, 2, \ldots, n-1\}$ be the random variable. Then

$$P_k = P\left[x = k\right] = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

So, the probability P_k is given by a binomial distribution, which is a bell shaped curve.[R84]

Average Shortest Path in a $G(n, p)$:

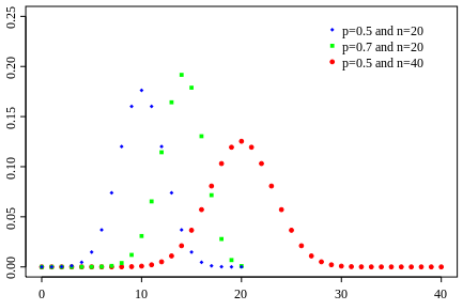

Figure 1.4: Binomial distribution as obtained in a G(n,p)

Given a graph $G(n, p)$,its degree distribution is binomial. Take a vertex $v \in G$ then expected number of neighbors of v are :

Let $k_v : V \longrightarrow \{1, 2, \ldots, (n-1)\}$ be the degree of v which is binomially distributed as discussed earlier so,

$$\begin{aligned} E[k_v] &= \sum_{i=0}^{n} i\binom{n}{i} p^i (1-p)^{(n-i)} \\ &= \sum_{i=1}^{n} i\binom{n}{i} p^i (1-p)^{(n-i)} \end{aligned}$$

use $i\binom{n}{i} = n\binom{n-1}{i-1}$

$$\begin{aligned} E[k_v] &= \sum_{i=0}^{n} n\binom{n-1}{i-1} p^i (1-p)^{(n-i)} \\ &= np \sum_{i=0}^{n} \binom{n-1}{i-1} p^{i-1} (1-p)^{(n-1)-(i-1)} \\ &= np \end{aligned}$$

Let $c = np$.
We know for some s,

$$\begin{aligned} c^s &= n \\ \implies slnc &= lnn \\ \implies s &= \frac{lnn}{lnc} \\ \implies s &= O(lnn) \end{aligned}$$

Let X be the average shortest path of a graph G
Now,

$$X = \left(1.c + 2.c^2 + 3.c^3 + ... + logn.c^{logn}\right) \frac{1}{\binom{n}{2}}$$

On solving the above AP-GP, we get that the average shortest path is of the $O\left(\frac{logn}{n}\right)$[**?**]

Scale-free Networks

Most of the systems in nature show a scale free behaviour, following a power law distribution. Albert and Barabasi have observed and concluded that many of the real-world networks follow two mechanism

- *Growth:* Starting with a small number (n_0) of vertices, at every time step a new node with $m(\leq n_0)$ edges to pre-existing nodes.
- *Preferential attachment:* The "Rich gets Richer phenomenon" i.e. a new edge tend to attach to a well connected node. In other words, the probability P that a new node is attached to a node i is proportional to the degree of the node i (k_i)

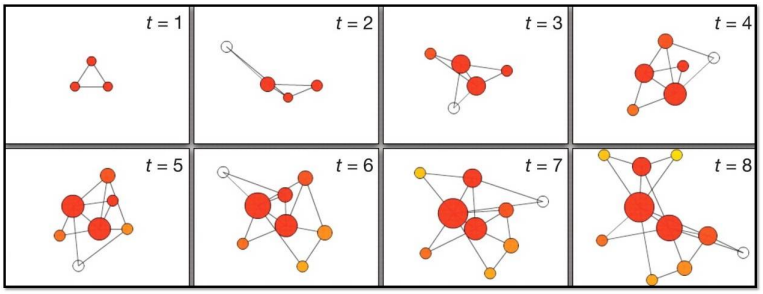

Figure 1.5: Scale free model www.sciencemag.org

$$P(k_i) = \frac{k_i}{\sum_j k_j}$$

After t time steps, there are $t + n_0$ nodes and mt edges.
Now, we shall try to derive the distribution of such networks, for that assume k_i is a continuous real variable. Then the rate at which k_i changes is proportional to the probability $P(k_i)$

$$\frac{dk_i}{dt} \propto P(k_i) \Rightarrow \frac{dk_i}{dt} = mP(k_i)$$

where m is the number of links added at each time step.
Now, substitute the value of $P(k_i)$

$$\frac{dk_i}{dt} = m\frac{k_i}{\sum_j k_j}$$

Since mt links are added at each time step t

$$\begin{aligned}\frac{dk_i}{dt} &= m\frac{k_i}{2mt} \\ &= \frac{k_i}{2t}\end{aligned}$$

Separating the variables, we have

$$\frac{dk_i}{k_i} = m\frac{dt}{2t}$$

Integrating both sides,

$$\ln k_i(t) = \frac{1}{2}\ln t + cont. \Rightarrow k_i = Ct^{1/2}$$

Let t_i be the time of inception of node i.
Then, we have the initial condition, $k_i(t_i) = m$, therefore,

$$k_i = m\left(\frac{t}{t_i}\right)^{1/2}$$

So, we have that the number of edges attached to the ith node varies as $t^{1/2}$.
Now, to find the cumulative probability distribution function,

$$\begin{aligned}P\left[k_i(t) < k\right] &= P\left[m\left(\frac{t}{t_i}\right)^{1/2} < k\right] \\ &= P\left(t_i > \frac{m^2 t}{k^2}\right)\end{aligned}$$

Assuming we add node at equal intervals to the system, the probability density of t_i is,

$$p(t_i) = \frac{1}{n_0 + t}$$

We know that there are $n_0 + t$ nodes and one node is added at each time step. Consequently,

$$\begin{aligned} P\left(t_i > \frac{m^2 t}{k^2}\right) &= \int_{\frac{m^2 t}{k^2}}^{n_0+t} p(t_i) dt_i \\ &= 1 - \int_0^{\frac{m^2 t}{k^2}} \frac{dt_i}{n_0 + t} \\ &= 1 - \frac{m^2 t}{k^2 (n_0 + t)} \end{aligned}$$

Therefore,

$$P\left[k_i(t) < k\right] = 1 - \frac{m^2 t}{k^2 (n_0 + t)}$$

Now, the probability density function $p(k)$,

$$\begin{aligned} p(k) &= \frac{dP[k_i(t) < k]}{dk} \\ &= \frac{2m^2 t}{n_0 + t} k^{-3} \end{aligned}$$

For $t \to \infty$, we have,

$$p(k) \sim 2m^2 k^{-3}$$

In such networks, the degree distribution is one-sided. It gives a power law form of distribution which is scale free as it is invariant under the change $k \longrightarrow bk$.

The World Wide Web is a common example of such networks, it has a few highly

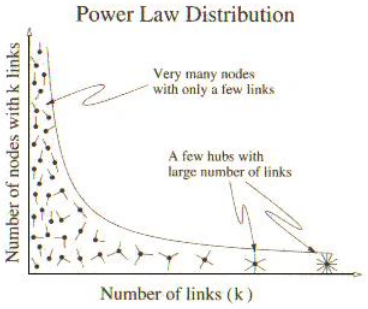

Figure 1.6: Scale free networks showing power law distribution

connected nodes(hubs) and a large number of spasely connected nodes, giving a fat-tail.

Small World

The *Small world phenomenon* could be described as "the principle that most of us are linked by short chain of acquaintances". The idea of the small world emerged through a real-world experiment in the late 1960s.This empirical study was undertaken by a social psychologist Stanley Milgram, who analyzed the average shortest path length for the social network of people of US(in which there was a link between two nodes if they know them on name basis). Milgram selected 296 "starter" individuals randomly from two US cities (Omaha, Nebraska and Wichita, Kansas), asking them to forward a letter to a "target" person in the town of Sharon, MA, a suburb in Boston. He gave them the target's name, address, occupations and some personal information. He asked these individuals to either forward this letter to the target person if they knew him or forward it to an acquaintance who was most likely to know the target person, with the goal of reaching the target as soon as possible. Roughly 20 percent of the letters arrived at the target with a median of six steps, proving the existence of short paths in a large network.[Mil67]
Milgram's experiment demonstrated the existence of abundant shortest paths in a large social network and the ability of individuals to effectively and collectively find these short paths without using a 'map' of the network. It is easy to imagine a social network with a lot of short paths but how does one choose the right path and how did a letter from thousands of miles reach the target getting from one acquaintance to another. It has been observed that the information about the target, his occupation and location were major factors in finding these short paths.
Further in this section we shall discuss the models based on the two principles of small worlds, the existence of short paths and the ability to find them. The existence of short paths is very intuitive. Suppose in a network every individual has atleast 100 friends. So, in two steps one could reach $100 \times 100 = 10,000$ people and in a matter of five steps one could reach 10 billion people which is approximately everyone in this world. Here the growth increases by a factor of 100 at each step, which does not hold true in real-world network as here we are assuming that every individual in a network knows 100 new people. In a real-world network we come across a lot of triadic closure i.e. three people who mutually know each other. With the effect of the triadic closure the number of people reached by the short paths is limited. This is what makes the small world phenomenon more surprising.

1. **Watts-Strogatz model[Wat98]:** This model exhibits both a network having closed triads and many short paths. Watts and Strogatz defined small world

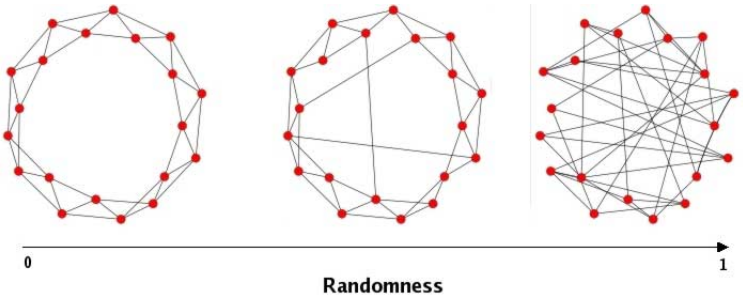

Figure 1.7: Small World Phenomenon[Wat98]

as a network having high clustering coefficient and low diameter. This model starts with a regular lattice(high clustering coefficient and high diameter) and progressively rewires the network (by picking an edge, disconnecting it from its end point and rewiring it to a random node) until we obtain a random graph(low clustering coefficient and small diameter). As the edges are rewired the diameter is reduced (farther the node being rewired lesser the diameter) and the clustering coefficient remains high.
Now, we shall analyze this model. Let R be a regular lattice with N sites and each vertex is connected to $2k$ nearest neighbours, k on each side. Each link connecting a node to its k nearest neighbour is rewired with a probability p and left the same with a probability $(1-p)$.

- *Impact on the diameter*:
  Let d_{ij} be the shortest distance between nodes i and j. Let $l(N,p)$ be the average shortest path defined as:

$$l(N,p) = \frac{1}{N(N-1)} \sum_i \sum_j d_{ij}$$

  Watts & Strogatz have shown that $l(N,p)$ decreases very rapidly with increasing $0 < p < 1$, approaching the limiting case. For $p = 0$, we have a linear chain of sites so $l(N,0) \sim N/4k$ and for limiting case $p \to 1$, the graph converges to a random graph with $l(N,1) \sim ln(N)/ln(2k-1)$.

- *Impact on Clustering Coefficient*:
  Let $C(p)$ be the clustering coefficient defined as

$$C(p) = \frac{1}{n} \sum_{i=1}^{n} C_i$$

where, for a node i with degree k_i and N_i as the set of neighbour we have,

$$C_i = \frac{2|e_{jk} : v_i, v_j \in N_i, e_{jk} \in E|}{k_i(k_i - 1)}$$

For a regular lattice $C(0) = \frac{3(k-2)}{4(k-1)}$ which tends to 3/4 as k grows independent of N and for a random graph $C(1) = \frac{K}{N}$. For $0 < p < 1$ the value of $C(p)$ remains quite close to $C(0)$ and falls for a relatively high p depicting the small world phenomenon.

2. **Kleinberg Model[Eas10]:** In the previous model, a contact probability was used to rewire the network. In Kleinberg's model the probability p to rewire the edges depends on the geographical distance between two nodes.
   Let N be the set nodes that are identified with a set of lattice points on a $n \times n$ square where $|N| = n$, $(i, j) : i \in 1, 2, \ldots, n, j \in 1, 2, \ldots, n$. Define lattice distance between two nodes $v_1 = (i, j)$ and $v_2 = (k, l)$ as $d(v_1, v_2) = |k - i| + |l - j|$.Each node in the network has an edge to all other nodes within lattice distance $p > 1$.Then we construct $q > 0$ directed edges from a node u to other nodes using independent random trials with a probability proportional to $[d(u, v)]^{-r}$ for $r > 0$.
   This model can be seen as the nodes as individuals living on the grid who know their neighbours for some number of steps in all direction and some acquaintances distributed broadly across the grid. Taking p and q as constants, we obtain a family by varying one parameter, r. For $r = 0$, we have the usual Watts & Strogratz model in which the rewiring is independent of the node's position on the grid. As r increases the long-range contacts of a node become more are more clustered in its vicinity.

1.2 Social Networks

Social network is a social structure which consists of individuals(social actors) and the ties between them. Networks have been studied in various different fields as objects of pure structure whose properties are fixed in time. This assumption doesn't hold good for real world networks,where the ties between individuals represent a communication, sending data, making decisions i.e. something they are doing and so it is not a fixed tie. A network can be viewed as a "continuously evolving and self constituting system", it is a dynamic system in which the structure of a node might affect the behaviour of

the whole system. Social network analysis(SNA) is a set of research procedures which views the social relationships in terms of network theory with the individuals as nodes and the relation between them as an edge. SNA studies the relations amongst nodes and the behaviour of the nodes.

1.2.1 Tools to Analyze

The position of an individual and the strength of its ties to other individuals is very important tools to analyze the network. The social position can be analyzed by the centrality measures and the strength of ties mostly involves closeness of bond.

Centrality Index

Centrality index can be defined as a real valued function on the vertex set $V(G)$ of a graph G. It is a structural index which means that if we have two isomorphic graph G and H and a mapping $\Phi : V(G) \longrightarrow V(H)$ then the centrality of $v \in V(G)$ is same as the $\Phi(v) \in V(H)$. In other words,Centrality index is a measure to find the relative importance of a vertex in a network. There are four major types of centrality measures: degree, betweenness,closeness and eigenvector.

1. *Degree Centrality:*
   Degree centrality $c_D(v)$ of a node v is defined as the degree of that node v, $deg(v)$ in case of an undirected graph. For a directed graph, we have the in-degree centrality $c_{iD}(v) = d^-(v)$ and out-degree centrality $c_{oD}(v) = d^+(v)$. Degree centrality is a local measure as it takes into account all the neighbours of a node.[Eas10]

2. *Betweenness Centrality:*
   Betweenness centrality quantifies the number of times a node acts as a bridge for the shortest path between a pair of nodes. Betweenness centrality is high for the nodes which have a high probability to occur on a randomly chosen shortest path between two randomly chosen nodes. Betweenness centrality of a graph G is defined as
   $$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$
   where $\sigma_{st}(v)$ is the number of shortest paths between nodes s and t passing through the node v and σ_{st} is the number of shortest paths between nodes s and t.[Eas10]

3. *Closeness Centrality:*

   In a graph distance is measured in terms of shortest path length.*Farness* of a node is defined as the sum of all the shortest path lengths from that node to all other nodes. Closeness centrality is the inverse of farness and for a node $v \in V$ it can be written as

   $$c_C(v) = \frac{1}{\sum_{u \in V} d(v,u)}$$

   Closeness centrality can be described as the time taken to spread the information from one node to all the other nodes sequentially.[Eas10]

4. *Eigenvector Centrality:*

   Eigenvector centrality of a node v assigns a relative score to the node v given the fact that connections to high scoring nodes will result in a higher score for v than the same number of connections to low-scoring nodes.
   For a graph $G(V,E)$ with adjacency matrix $A = (a_{ij})$ we define eigenvector centrality as

   $$c_E(v) = \frac{1}{\lambda} \sum_{t \in N(v)} c_E(t) = \frac{1}{\lambda} \sum_{t \in V} a_{vt} c_E(t)$$

   where $N(v)$ is the set of neighbours of the node v and λ is a constant. We can write this as

   $$Ax = \lambda x$$

   There might be different values of λ but an eigenvector with all entries positive is given by the highest eigenvalue. Then vth component of the related eigenvector gives the centrality score.[Eas10]

These measures give the importance of a node in a network locally and globally.

Strength of Weak Ties

Strength of weak ties is a concept in social network analysis which was first proposed by a sociologist, Mark Granovetter.He defined strength of a tie as a linear combination of

- F:Frequency of contact
- E:Emotional Intensity
- I:Intimacy (mutual confiding)

- R:Reciprocal services

$$ST_{ij} = w_1F + w_2E + w_3I + w_4R$$

Granovetter used an empirical example to describe his concept, he did a survey of job seekers. He asked the ones who had found a job through a contact that how often they meet those people and how well they know their contact. Around 60 percent saw their contact occasionally and 30 percent rarely knew them. Then he argued that if A is friends with B and C then with a high probability B knows C and if A needs a job then his friends B and C will have same kind of information as A. So, an acquaintance like D, weak tie is more probable to have some new information. The more the social distance between A and D more beneficial is the weak tie for A.

Granovetter also argues that a strong tie can never be a bridge but weak ties act

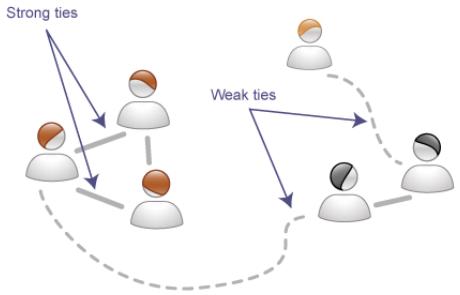

Figure 1.8: Strong and Weak ties

as bridges connecting disconnected social groups. This is because if someone has a strong tie with someone else then the people around them are most likely to be tied to them. These weak ties play an important role in diffusion in a network. [Gra73]

1.3 Conclusion

In this chapter we discussed the bascis of network theory and the ways to analyze it.We discussed communities in networks, its importance and the methods to detect it. In the next chapter we shall build up a model based on disconnected communities in a real-world network and study the dynamics of the average distance between these communities when edges are added between them. We will also try to justify the concept of strength of weak ties through our model.

Chapter 2

2.1 Motivation

The Indian society has always been intrinsically linked to the caste system.Caste is defined to be a Hindu hereditary class of socially equal persons, united in religion and usually following similar occupations, distinguished from other castes in the hierarchy by its relative degree of purity or pollution. Caste system is said to have originated with the Vedas, the mythological text. There were four different castes, The *Brahmins* (scholar or priests),the *Kshatrias* (soldier class),the *Vaishias* (business men),and the *Sudras* (menial working class). The occupational roles were determined according to the castes and as time progressed this led to the association of castes with economic status and hence the social status. With such practices, the caste based discrimination became predominant in the society and created a rift between the upper castes (higher economic status) and the lower castes(low economic status) and divided the society into two disconnected clusters. The upper caste or the forward classes progressed year by year but the lower castes or the backward classes remained the same. To offset the practice of discriminatory social stratification based on castes,affirmative action steps were undertaken to uplift the backward classes.In this chapter, we shall discuss this system from a network theoretic perspective. We shall study the two social communities, the *socially forward and uplifted* (FC) and the *socially backward and downtrodden* (BC).

Our motivation comes from the existence of a tangible strength associated with every weak tie, as proposed by Mark S. Granovetter in his famously cited theory of the Strength of Weak Ties(discussed in the previous chapter). Past studies in network analytics by Matthew O. Jackson has shown that network formation and subsequent interaction between the nodes is highly influenced by homophily. In India, associations amongst the people is seen to be largely determined by caste-based homophily,

hence for the purposes of our study, we choose to term the network formation pattern among the Indian population as a caste-based homophilic network.
We establish why and how the reservation system maintains a very good balance between the two. We shall discuss a mathematical model which will help us quantify the reservation procedure and show how the reservation system is affecting the socially backward classes.Here we shall consider the number of links between the two communities as a measure of stability of the large scale social structure. We quantify the *forward breeze effect*, which is, to put simply, the change in mind-set of the FC, on being in contact with the BC. We also quantify the *reverse breeze effect*, which is the increased motivation felt by the BC to achieve upliftment, by being influenced by those close to them and around them. We try to find the optimum number of links required between the two communities.

2.2 Model Network Structure

In this section, we shall discuss the structure of our network model. The structure of the network is dynamic and is modelled by a finite undirected graph defined as $G = (V, E)$ where V is a finite set of nodes and E is the set of edges. Each node $v \in V$ represents an individual in a community, while each undirected edge $e = (u, v) \in E$ represents a tie between two individuals.

We initially define G to consist of two symmetric, disconnected clusters of nodes representing the two communities under consideration: the socially forward and uplifted community and the socially backward and downtrodden community. We note that due to a history of social segregation and caste-based homophily, there are few interactions or links between members of different social groups. Each cluster is generated as an individual *Erdös-Renyi* graph $G_{ER} = (n, p')$, where edges between any two of the n nodes are added with probability p' independent from every other edge. As discussed in the previous chapter,in an Erdös-Renyi graph the total number of edges is given by $\frac{n(n-1)p'}{2}$
We define the socially uplifted community cluster as $FC \mid FC \subset V$ and the socially backward community cluster as $BC \mid BC \subset V$ such that $FC \cap BC = \phi$, the number of nodes $n_G = |V|$, the total number of nodes connected from BC to FC as n_{inter}

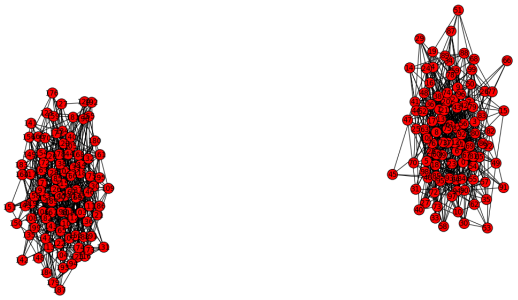

Figure 2.1: Disconnected Communities

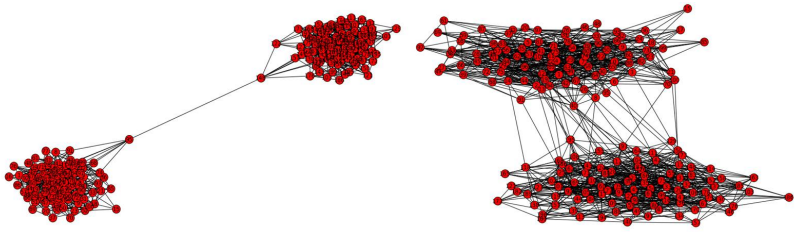

Figure 2.2: Bridges added between disconnected communities

2.3 Empirical Results

First let us take two clusters, and add edges between two randomly selected nodes and then calculate the average shortest path of the graph G using the formula:

$$\text{Average Shortest Path(avg)} = \sum_{i,j} d(v_i, v_j) \frac{1}{n(n-1)}$$

The graph obtained:

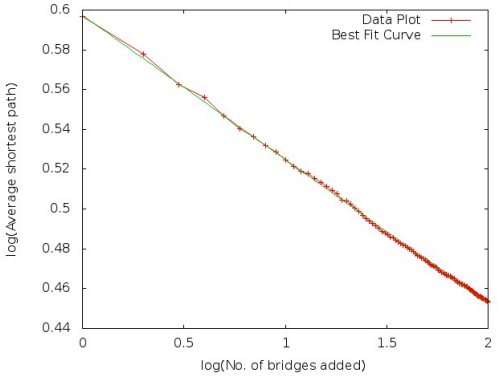

Figure 2.3: log-log plot

In the plot of $log(avg)$ vs $log(n_{inter})$ we get a straight line. From the graph we can

conclude that the graph of avg vs n_{inter} could follow a power law i.e. have a fat-tailed distribution.
Now, we shall calculate the average shortest path between the clusters FC and BC using the formula:

$$\text{Average Shortest Path(avgc)} = \sum_{i \in G_1, j \in G_2} d(v_i, v_j) \frac{1}{n^2}$$

The graph obtained:

From the graph in figure we observe that the average shortest path between the

[h!]

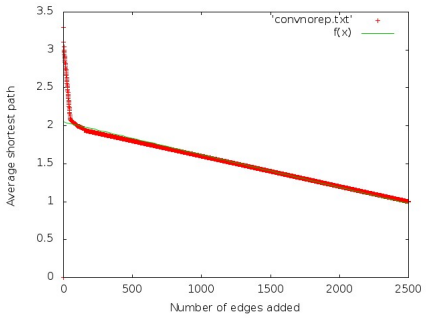

Figure 2.4: Convergence plot

clusters drop to a very small value of 2 in a small interval of time.We can see from the plot that avgc drops to 2 when n_{inter} is around 100, which is a very small number compared to possible number of edges between the clusters which is 2500.
With the above two empirical results we can conclude that in order to get a small value for the spread of influence or the average shortest path we need a very small number of bridges between the clusters, and we can see that after this value if more bridges are added, it will not make a lot of difference to the average shortest path.

2.4 A Different Perspective

To explain our empirical observations and the model,we define an analogy for the same in this section. Now, our problem is defined as:

- Let the two disconnected communities be a random graph.
- Now, look at the problem as if the nodes of the network were random points on a square on the Euclidean plane.

- Select some k pair of points at random on the plane, (X_i, Y_i) such that if one reaches X_i it gets teleported to Y_i.

- Define distance between two points as

$$Dis(A, B) = \min_i [d(A, X_i) + d(B, Y_i)] \tag{2.1}$$

  where $d(x_1, x_2)$ is the Euclidean distance between the points x_1 and x_2.

- Then what is the expected distance between any two points in this square on the plane.

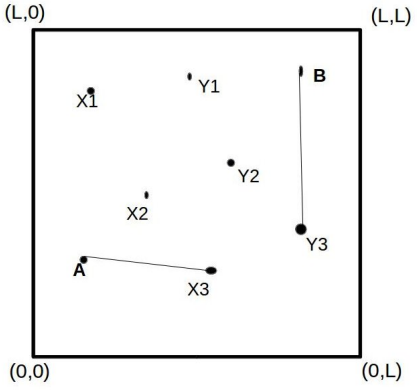

Figure 2.5: Convergence plot

2.4.1 Theoretical Observations

Let (X_1, Y_1) and (X_2, Y_2) be two random points. Assume that X_1 and X_2 are independent and evenly distributed on an interval $(0, n)$ and same for Y_1 and Y_2.[Phi07]

One Dimensional

First let us calculate the probability distribution function $F(t) = P((X_1 - X_2)^2 \leq t)$

$$\begin{aligned}
P\left[(X_1 - X_2)^2 \leq t\right] &= 1 - P\left[(X_1 - X_2)^2 \geq t\right] \\
&= 1 - P\left[(X_1 - X_2) \geq \pm\sqrt{t}\right] \\
&= 1 - P\left[\left((X_1 - X_2) \geq \sqrt{t}\right) \bigcap \left((X_1 - X_2) \geq -\sqrt{t}\right)\right] \\
&= 1 - \left[P\left[(X_1 - X_2) \geq \sqrt{t}\right]\right]^2 \\
&= 1 - \left[1 - P\left[(X_1 - X_2) \leq \sqrt{t}\right]\right]^2 \\
&= 1 - \left[1 - \frac{\sqrt{t}}{n}\right]^2
\end{aligned}$$

So, we have

$$F(t) = \begin{cases} 1 - \left[1 - \frac{\sqrt{t}}{n}\right]^2 & 0 < t < a^2 \\ 1 & a^2 < t \end{cases}$$

Now the corresponding density function is

$$f(t) = \frac{dF(t)}{dt}$$

$$f(t) = \begin{cases} \frac{1}{n\sqrt{t}} - \frac{1}{n^2} & 0 < t < n^2 \\ 0 & \text{otherwise} \end{cases}$$

Two Dimensional (Square)

Now, we will calculate the probability density function of the event $\left[(X_1 - X_2)^2 + (Y_1 - Y_2)^2 \leq s\right]$ we shall take the convolution g of f and f

$$g(s) = \int f(s-t)f(t)dt$$

$$g(s) = \begin{cases} -4\frac{\sqrt{s}}{n^3} + \frac{\pi}{n^2} + \frac{s}{a^4} & 0 < s < n^2 \\ -\frac{\pi}{n^2} - \frac{s}{a^4} & n^2 < s < 2n^2 \end{cases}$$

Expected Distance in a Square

The expected distance between two points in a square is

$$E_{square} = \int_0^{2n^2} \sqrt{s} g(s) ds = \int_0^{\sqrt{2}n} v g_v(v) dv$$

where $v = \sqrt{s}$ and $g_v(v) = g(v^2)\frac{ds}{dv} = 2vg(v^2)$ So,

$$E_{square} = \frac{n}{3}\ln(1+\sqrt{2}) + \frac{1}{15}(2n + n\sqrt{2})$$

2.4.2 Comparison

In this section we shall compare our initial problem and the analogy by sampling data. The plots obtained by sampling are

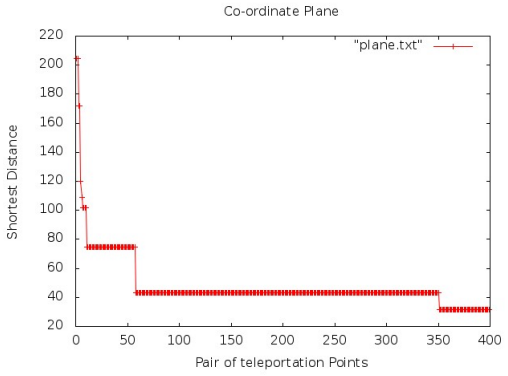

Figure 2.6: Shortest distance vs No. of teleportation points

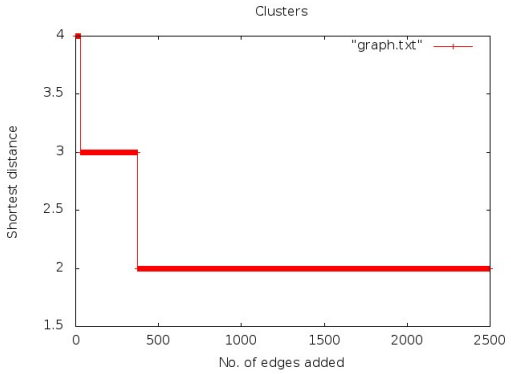

Figure 2.7: Shortest distance vs No. of bridges added

- Figure 2.6 shows a graph where we have points on a square in the Euclidean plane and the plot shows the variation of the shortest distance between two fixed points vs the number of teleportation points.
- Figure 2.7 shows a graph where we have two disconnected clusters and the plot shows the variation of the shortest distance between two fixed nodes vs the number of bridges between the clusters.
- Looking at both the graphs we observe that both the plots show a similar variation. In fig 2.7 the shortest distance becomes 2 at around when 400 edges are added and similar kind of drop is observed in figure 2.6.

2.5 Solution

In this section we shall calculate the distribution of the distance between two random points[Mol12] and hence the expected distance between two points on a square on a Euclidean plane using (2.1).
Consider two random points $A_1 = (x_1, y_1)$ and $A_2 = (x_2, y_2)$ on a square of side L on the co-ordinate plane (as shown in the figure) where x_1, x_2, y_1, y_2 are identically uniformly distributed over $[0, L]$.
Now, $\Delta_X = (x_1 - x_2)$ is also a random variable. We want to find its probability density, $f(\Delta_X)$.
We can write $x_1 - x_2$ as $x_1 + (-x_2)$ i.e. the sum of independent random variables.

Theorem 2. *Let X and Y be two independent random variables with density $h_X(x)$ and $h_Y(y)$ defined for all x and y. Then the sum $Z = X + Y$ is also a random variable with density $h_Z(z)$ which is calculated by the convolution of f_X and f_Y defined as*

$$\begin{aligned} h_Z(z) &= \int_{-\infty}^{\infty} h_X(z - x) h_Y(x) dx \\ &= \int_{-\infty}^{\infty} h_X(z - y) h_Y(y) dx \end{aligned}$$

So, we can calculate $f(\Delta_X)$ as

$$f(\Delta_X) = \int_{-\infty}^{\infty} f(x) f(x - \Delta_X) dx \tag{2.2}$$

where $f(x) = 1/L$ as x is uniformly distributed.
We observe that $f(x) \neq 0$ for $x \in [0, L]$ and so $f(x - \Delta_X) \neq 0$ for $0 \leq (x - \Delta_X) \leq L$. Thus we need to solve (2.1) for $(-L, \Delta_X]$ when $\Delta_X \leq 0$ and $(\Delta_X, L]$ when $\Delta_X > 0$. After solving (2.1), we have

$$f(\Delta_X) = \begin{cases} 0 & \Delta_X \notin (-L, L] \\ \frac{\Delta_X + L}{L^2} & \Delta_X \in (-L, 0] \\ \frac{L - \Delta_X}{L^2} & \Delta_X \in (0, L] \end{cases} \tag{2.3}$$

Now, we need the probability distribution function of Δ_X^2 so we shall use the following theorem,

Theorem 3. *Let X be a random variable with probability density function $f(x)$ and let $Y = \phi(x)$ be another random variable with probability density function $g(y)$ and cumulative distribution function $G(y)$. If $\phi(x)$ is monotonous and differentiable in $[a, b]$ then $g(y)$ and $G(y)$ are given by*

$$g(y) = \begin{cases} f(\phi(y)) \dfrac{d\phi(y)}{dy} & \dfrac{d\phi(x)}{dx} > 0 \forall x \in [a, b] \\ -f(\phi(y)) \dfrac{d\phi(y)}{dy} & \dfrac{d\phi(x)}{dx} < 0 \forall x \in [a, b] \end{cases} \tag{2.4}$$

$$G(y) = \begin{cases} \int_a^{\phi(y)} f(x) dx & \dfrac{d\phi(x)}{dx} > 0 \forall x \in [a, b] \\ \int_{\phi(y)}^b f(x) dx & \dfrac{d\phi(x)}{dx} < 0 \forall x \in [a, b] \end{cases} \tag{2.5}$$

where conditions imply that a function is either increasing or decreasing in $[a, b]$

Observe that (2.2) is monotonously increasing in $\Delta_X \in (-L, 0]$ and monotonously decreasing in $\Delta_X \in (0, L]$. We know that $Pr\{\Delta_X^2 \leq 0\} = 0$ and $Pr\{\Delta_X^2 \geq L^2\} = 0$.So, we have,

$$g(y) = \begin{cases} 0 & \Delta_X^2 \notin (0, L^2] \\ \frac{1}{L\sqrt{\Delta_X}} - \frac{1}{L^2} & \Delta_X^2 \in (0, L^2] \end{cases} \tag{2.6}$$

Next, we shall determine the distribution of $u = (\Delta_X^2 + \Delta_Y^2)$. Observe that Δ_X^2 and Δ_Y^2 are independent random variables and their distribution is same. Thus the

cumulative distribution function of u can be written as

$$F_U(u) = \int\int_A f(\Delta_X^2) f(\Delta_Y^2) d\Delta_X^2 d\Delta_Y^2 \tag{2.7}$$

where A is the area below $u = (\Delta_X^2 + \Delta_Y^2)$ inside the square of size $L^2 \times L^2$.
On solving (2.6) we get,

$$F_U(u) = \begin{cases} \pi a - \frac{8}{3}a^{3/2} + \frac{a^2}{2} & u \in [0, L^2) \\ 1 - \left(\frac{2}{3} + 2a + \frac{a^2}{2} - \frac{2(a-1)^{3/2}}{3} - 2\sqrt{a-1} - 2a\sqrt{a-1} - 2a \arcsin(\left(\frac{2-a}{a}\right)\right) & u \in [L^2, 2L^2) \\ 0 & u \in (-\infty, 0) \\ 1 & u \in (2L^2, \infty) \end{cases} \tag{2.8}$$

Finally, we will calculate $F_K(l)$,$l = \sqrt{u} = \sqrt{\Delta_X^2 + \Delta_Y^2}$

$$F_K(l) = Pr\left\{K < l\right\} = Pr\left\{0 < u < l^2\right\} \ , \ l \in (0, L\sqrt{2}) \tag{2.9}$$

and

$$Pr\left\{0 < u < l^2\right\} = \int_0^{l^2} f(u) du \tag{2.10}$$

Using (2.3) we get

$$F_K(l) = \begin{cases} \frac{\pi l^2}{L^2} - \frac{8l^3}{3L^3} + \frac{l^4}{2L^4} & l \in [0, L) \\ 1 - \left(\frac{2}{3} + 2b + \frac{b^2}{2} - \frac{2(b-1)^{3/2}}{3} - 2\sqrt{b-1} - 2b\sqrt{b-1} - 2b \arcsin\frac{2-b}{b}\right) & l \in [L, \sqrt{L}) \\ 0 & u \in (-\infty, 0) \\ 1 & u \in (\sqrt{L}, \infty) \end{cases} \tag{2.11}$$

where $b = \frac{l^2}{L^2}$

Now, we shall calculate the expected minimum distance between two points.

Theorem 4. *Let $Y_1, Y_2, \ldots, Y_n$ be independent real-valued random variables with same distribution say, F.Let Y_k' be the kth smallest variable of $(Y_1, Y_2, \ldots, Y_n)$ i.e. say $(Y_1', Y_2', \ldots, Y_n')$ is a permutation of $(Y_1, Y_2, \ldots, Y_n)$ such that $Y_1' \leq Y_2' \leq \cdots \leq Y_n'$).Then the distribution function of Y_k' is given by*

$$G_k = \sum_{w=k}^{n} \binom{n}{w} (F)^w (1-F)^{n-w}, \ k = 1, \ldots, n \tag{2.12}$$

Applying the above theorem to $F_K(l)$ we will get the distribution of the minimum distance between two random points and hence the expectation.

2.6 Conclusion

In this chapter we discussed the network model based on the reservation system in India. We observed that on putting a few number of edges between the disconnected communities the average shortest path decreases drastically. This shows that the strength of weak ties play a role here as we introduced a few strong ties it resulted in a large number of weak ties and the average shortest path reduced. We also observe that once the average shortest path decreases to this number then it decreases very slowly.

In terms of the reservation system, we can conclude that this system is surely bridging the gap between the two disconnected communities. In future we would like to investigate the question that what is the optimum number of bridges to be put in the network such that there is no significant change in the average shortest path.

Bibliography

[APL90] Horst D. Simon Alex Pothen and Kan-Pu Liou, *Partitioning sparse matrices with eigenvectors of graphs.*, Society for Industrial and Applied Mathematics (SIAM) Journal on Matrix Analysis and Applications **11** (1990), 430–452.

[Bol98] Bëla Bollobäs, *Modern graph theory*, Springer-Verlag New York, 1998.

[Eas10] Kleinberg Jon Easley, David, *Networks, crowds and markets*, Cambridge University Press, 2010.

[FRP04] Federico Cecconi Vittorio Loreto Filippo Radicchi, Claudio Castellano and Domenico Parisi., *Defining and identifying communities in networks*, Proceedings of the National Academy of Sciences (PNAS) **101** (2004), 2658–2663.

[Gra73] Mark Granovetter, *The strength of weak ties*, American Journal of Sociology **78** (1973), 1360–1380.

[J.03] Watts Duncan J., *Six degrees: The science of a connected age*, W.W.Norton & Company, February 2003.

[KL70] Brian W. Kernighan and Shen Lin, *An efficient heuristic procedure for partitioning graphs.*, Bell System Technical Journal **49** (1970), 291–307.

[M.E04] Newman M.E.J, *Detecting community structure in networks*, The European Physics Journal B **38** (2004), 321–330.

[M.F73] M.Fiedler, *Algebraic connectivity of graphs*, Czech. Math. J. **23** (1973), 298–305.

[Mil67] Stanley Milgram, *The small world problem*, Psychology Today **1** (1967).

[Mol12] D. Moltchanov, *Survey paper: Distance distributions in random networks*, Ad Hoc Networks **10** (2012), 1146–1166.

[New06] M.E.J Newman, *Modularity and community structure in networks*, Proceedings of the National Academy of Sciences (PNAS) **103** (2006), 8577–8582.

[Phi07] J. Philip., *The probability distribution of the distance between two random points in a box.*, TRITA MAT **7** (2007).

[R84] P. Erdős & and A. Rényi, *The evolution of random graphs*, Phys. Rev. Lett. **286** (1984), 343–347.

[Wat98] S. H. Watts, D. J. & Strogatz, *Collective dynamics of small world*, Nature **393** (1998), 440–442.