Compressed Sensing Over Graphs

Identifying central nodes for information flow by estimating betweenness centrality of target nodes

Group Members

Name	Roll Number	Contact
Aaron Jerry Ninan	190100001	190100001@iitb.ac.in
Richeek Das	190260036	richeek@cse.iitb.ac.in

Papers and References

This paper [Compressed Sensing Over Graphs] stood out as one of the base papers, which defined the idea formally.

The following paper defines the problem of identifying nodes with high betweenness centrality. It also proposes an algorithm "DICeNod" for solving the same. **[We implement this algorithm in Python and check for the viability of the proposed results.]**

- Identifying central nodes for information flow in social networks using compressive sensing.

Datasets Used

- **Stanford Large Network Dataset Collection** serve as our starting datasets for experimentation. Link - SNAP Datasets.

  For our experiments we focus on two datasets,

  (a) **Facebook:** This is an anonymized dataset consists of 'circles' (or 'friends lists') from Facebook. Facebook data was collected from survey participants using this Facebook app. The dataset includes node features (profiles) and circles. **[(4039,88234)]** nodes and edges.

  (b) **GNUTELLA:** This is a snapshot of the Gnutella peer-to-peer file sharing network from August 8, 2002. Nodes represent hosts in the Gnutella network topology and edges represent connections between the Gnutella hosts. **[(6301,20777)]** nodes and edges.

Motivation

Betweeness centrality is one of the prominent measures which shows the node importance from the information flow status in the network. Therefore, we are interested in finding the nodes with high centrality. If we take an example of companies wanting to give free sample products to social media influencers for reviewing/advertising, they must choose major hotspots of information flow! Identifying these nodes as we will see later can be a very conservative problem.

Problem Statement

Given end to end measurements in a network with unknown topology we are required to reconstruct the vector containing local centralities of the nodes of the graph.

Betweenness Centrality:

$$C_B(u) = \sum_{v,w,v\neq w} \frac{\sigma_{vw}(u)}{\sigma_{vw}}$$

where σ_{vw} is the total number of shortest paths between v and w. $\sigma_{vw}(u)$ is the number of such paths that pass through u.

Our assumption: Very few nodes have high centrality values, this makes the vector we want to reconstruct to be sparse.

Measurement Matrix Design

We choose some parameters and build a measurement matrix. We will later see that we get some theoretical guarantees if our measurement matrix follows these guidelines.

$$\epsilon \in (0, \frac{1}{6}), \theta \in [0, 1), \mu > 0, C > 1$$

$$d = \frac{1}{\epsilon} \frac{\log(\frac{e(\theta+1)n}{\theta k'})}{\mu C}$$

$$m = \frac{Cdk'}{\epsilon}$$

Algorithm 1: MeasurementMatrix

Input: d,m,n
Output: A

```
A_{m×n} = 0_{m×n}
for j = 1 to n do
    row[] = Select randomly d numbers in 1,2,···,m
    for i ∈ row do
        A[row[i],j] = 1
    end
end
```

Proof of Theoretical Guarantees

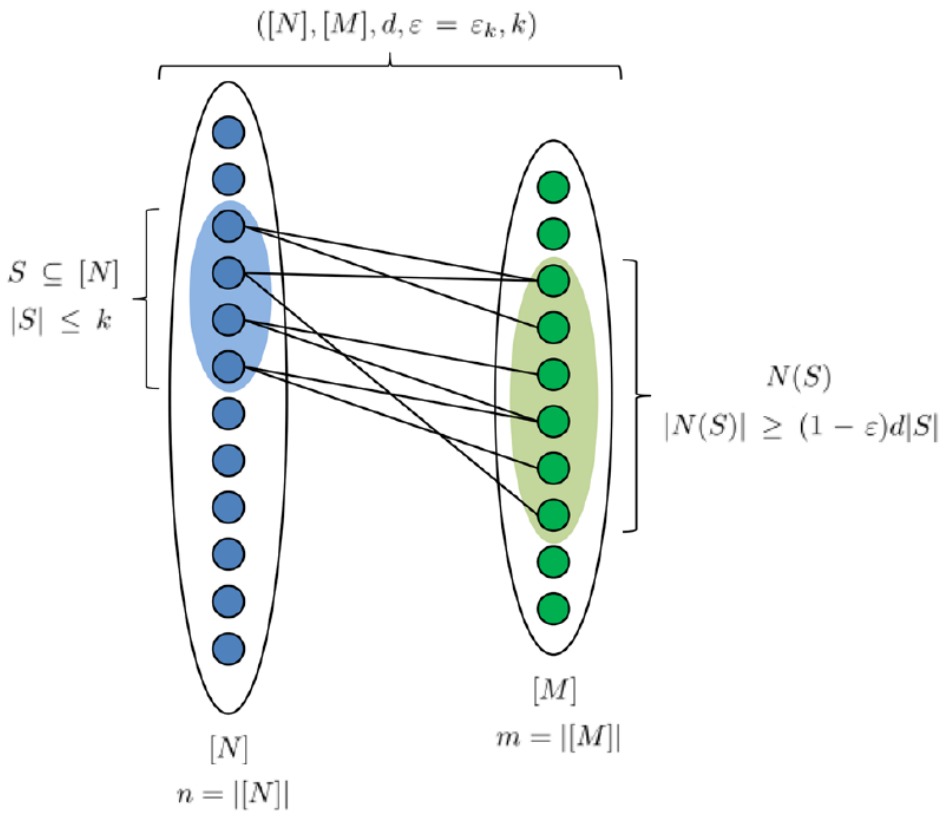
Definition 1: Let $([N], [M], d, \epsilon = \epsilon_k, k)$ be a left-regular bipartite graph with $n = |[N]|$ left nodes, $m = |[M]|$ right nodes, a set of edges ζ and a left degree d. If for some $\epsilon \in (0, \frac{1}{6})$ and any $S \subset [N]$ of size $|S| \leq k$ we have $|\mathcal{N}(S)| \geq (1-\epsilon)d|S|$, then this graph is referred to as **lossless bipartite expander graph**.

$$\text{Here, } |\mathcal{N}(S)| = \{j \in [M] | \exists i \in S, e_{ij} \in \zeta\}$$

The constructed measurement matrix $\mathcal{A}$ corresponds to bi-adjacency matrix of a left-regular bipartite graph G

Now,we want to find lower bounds for $m = [M]$ and d such that d-left-regular bipartite graph would be a lossless expander graph with very high probability.

Example of a lossless bipartite expander graph with $k = 4$ and $d = 2$



Let π indicate the event of G not being a $([N], [M], d, \epsilon, k)$-lossless expander graph.

$$\pi = \bigcup_{s=1}^{k} \pi_s$$

$$\pi_s = \bigcup_{S \subseteq N, s=|S|} [|\mathcal{N}(S)| < (1 - \epsilon)ds]$$

Apply Union Bound Theorem to both equations above, we have,

$$Pr(\pi) = \sum_{s=1}^{k} Pr(\pi_s)$$

$$Pr(\pi_s) = \sum_{S \subseteq N, s=|S|} Pr[|\mathcal{N}(S)| < (1 - \epsilon)ds]$$

Our task is to find an upper bound for $Pr(\pi)$.

For this we only need to find an upper bound to $Pr[|\mathcal{N}(S)| < (1 - \epsilon)ds]$

We utilize a variant of **Chernoff bound** introduced in the following lemma.

Lemma There exists constants $\mu > 0$ and $C > 1$ such that if $m \geq \frac{Cds}{\epsilon}$, for any $S \subseteq [N]$ with $|S| = s$, then one has:

$$Pr[|\mathcal{N}(S)| < (1-\epsilon)ds] \leq (\mu\frac{\epsilon m}{ds})^{-\epsilon ds}$$

Using this we get,

$$Pr(\pi_s) \leq \sum_{S\subseteq N, s=|S|} (\mu\frac{\epsilon m}{ds})^{-\epsilon ds}$$
$$\implies Pr(\pi_s) \leq \binom{n}{s}(\mu\frac{\epsilon m}{ds})^{-\epsilon ds}$$
$$\implies Pr(\pi) \leq \sum_{s=1}^{k}\binom{n}{s}(\mu\frac{\epsilon m}{ds})^{-\epsilon ds}$$

Put $\gamma = \frac{k}{n}$, the above inequality after **Sterling Approximation** becomes,

$$Pr(\pi) \leq \sum_{s=1}^{\gamma n}(\frac{ne}{s})^{s}(\mu\frac{\epsilon m}{ds})^{-\epsilon ds}$$
$$\text{Using, } m \geq \frac{Cds}{\epsilon} \text{ and } s \leq \gamma n, \text{ we have}$$
$$Pr(\pi) \leq \sum_{s=1}^{\gamma n}(\frac{e}{\gamma}(\mu C)^{-\epsilon d})^{s}$$
$$\text{Let } x = \frac{e}{\gamma}(\mu C)^{-\epsilon d}$$
$$\implies Pr(\pi) < \sum_{s=1}^{\infty} x^s = \frac{x}{1-x}$$

x can be chosen in a way that the term on the right-hand side can be an arbitrarily small constant θ ($Pr(\pi) < \theta$)

Hence we have,

$$x = \frac{\theta}{\theta+1}$$
$$\implies d = \frac{1}{\epsilon}\log_{\mu C}\frac{e(\theta+1)n}{k\theta}$$

Hence we obtain a d-left regular lossless bipartite expander graph with probability at least $1-\theta$ with $m = \frac{Cdk}{\epsilon}$ right nodes and n left nodes, where $d = \frac{1}{\epsilon}\log_{\mu C}\frac{e(\theta+1)n}{k\theta}$

How do we ensure that the measurement matrix formed this way will induce a Sub-graph in a row of measurement?

We prove that this method ensures that a connected sub-graph is induced almost always with high probability.

For a given row r in matrix A, corresponding to a measurement, let $X_{r[i]}$ be the random indicator variable that is set to 1 if the node i is visited by that measurement, and set to 0 otherwise. As we choose d indices from a set of measurements with cardinality m, uniformly at random for each column of A , the probability of an arbitrary node for being visited in that given measurement is $\frac{d}{m}$. We define X_r as the random variable that indicates the number of visited nodes in that given measurement.

$$X_r = \sum_{i=1}^{n} X_{r[i]}$$
$$\delta = E[X_r] = \sum_{i=1}^{n} E[X_{r[i]}] = \sum_{i=1}^{n} \frac{d}{m} = \frac{nd}{m}$$

Theorem:-

Let $X_{r[1]}, X_{r[2]}, X_{r[3]}, ...$,be indpendent random variables, not necessarily with the same distribution and $0 \leq X_{r[i]} \leq 1$. Let $X_r = X_{r[1]} + X_{r[2]} + ... + X_{r[n]}$, then for any $\epsilon', \epsilon'' \geq 0$, we have

$$Pr[X_r \geq (1+\epsilon')\delta] \leq e^{-\frac{\epsilon'^2}{\epsilon'+2}\delta}$$
$$Pr[X_r \leq (1-\epsilon'')\delta] \leq e^{-\frac{\epsilon''^2}{2}\delta}$$

By putting suitable values of ϵ', ϵ'' and $\gamma \to 0$ we can obtain,

$$Pr[X_r \geq (1+\epsilon')\delta] \leq e^{-\frac{1}{\gamma}} \to 0$$
$$Pr[X_r \leq (1-\epsilon'')\delta] \leq e^{-\frac{\epsilon}{4C\gamma}} \to 0$$

This proves that almost always the sequence of visited nodes in each measurement of $\mathcal{A}$ is no longer than $L_{max} = (1+\epsilon')\frac{\epsilon}{C\gamma}$ and not less than $L_{min} = (1-\epsilon'')\frac{\epsilon}{C\gamma}$

Erdős–Rényi model for Random Graphs:-

In the $G(n,p)$ model, a graph is constructed by connecting labeled nodes randomly. Each edge is included in the graph with probability p, independently from every other edge. The parameter p in this model can be thought of as a weighting function; as p increases from 0 to 1, the model becomes more and more likely to include graphs with more edges and less and less likely to include graphs with fewer edges.

Theorem:-

A general graph in the G(n, p) model is almost surely connected if $p > \frac{\ln n}{n}$

Using this we get, An induced subgraph over any arbitrary row of the measurement matrix $\mathcal{A}$ in DICeNod is almost surely connected, when it corresponds to a subgraph of the generalized $G(n,p)$ graph model with $p > \frac{\ln(L_{min})}{L_{min}}$

As the networks we are considering are very large, hence $\gamma \to 0 \implies L_{min} \to \infty$.

$$\implies \frac{\ln(L_{min})}{L_{min}} \to 0$$
$$\implies p > 0$$

Hence, the only requirement for our sub-graph to be connected is $p > 0$, which is trivially true !!

Recovery Guarantees:-

We have an equation of the form,

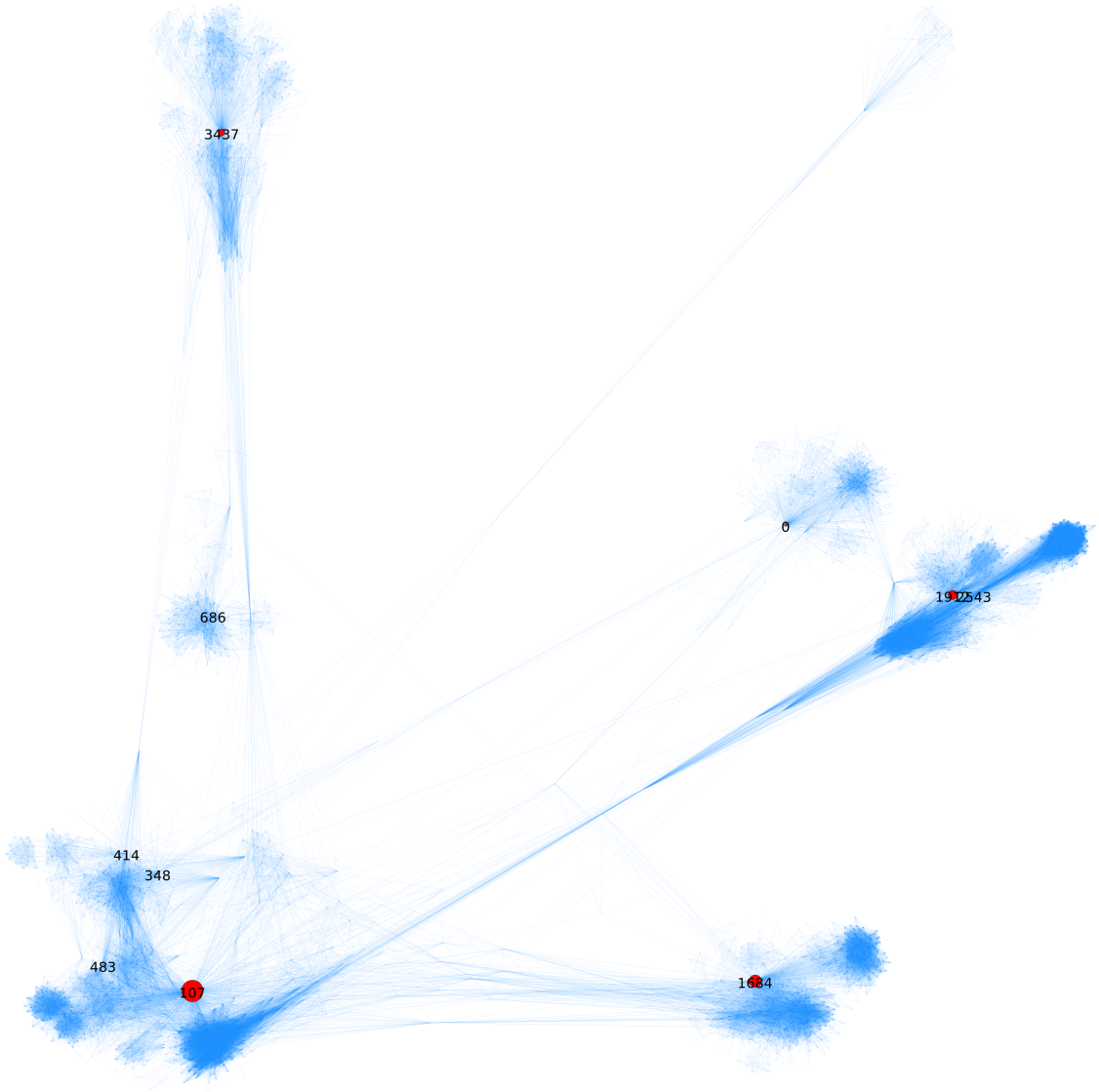
$$||\hat{x} - x^*||_1 \leq c||x^* - x_S^*||_1$$
$$\text{such that, } c = \frac{2}{1 - \frac{4\epsilon}{1-2\epsilon}}$$

Note that, $c > 0 \implies \epsilon < \frac{1}{6}$, which we have already considered in construction of measurement matrix A.

The above result is taken from the paper [Combining geometry and combinatorics: A unified approach to sparse signal recovery]

Results/Working of Code

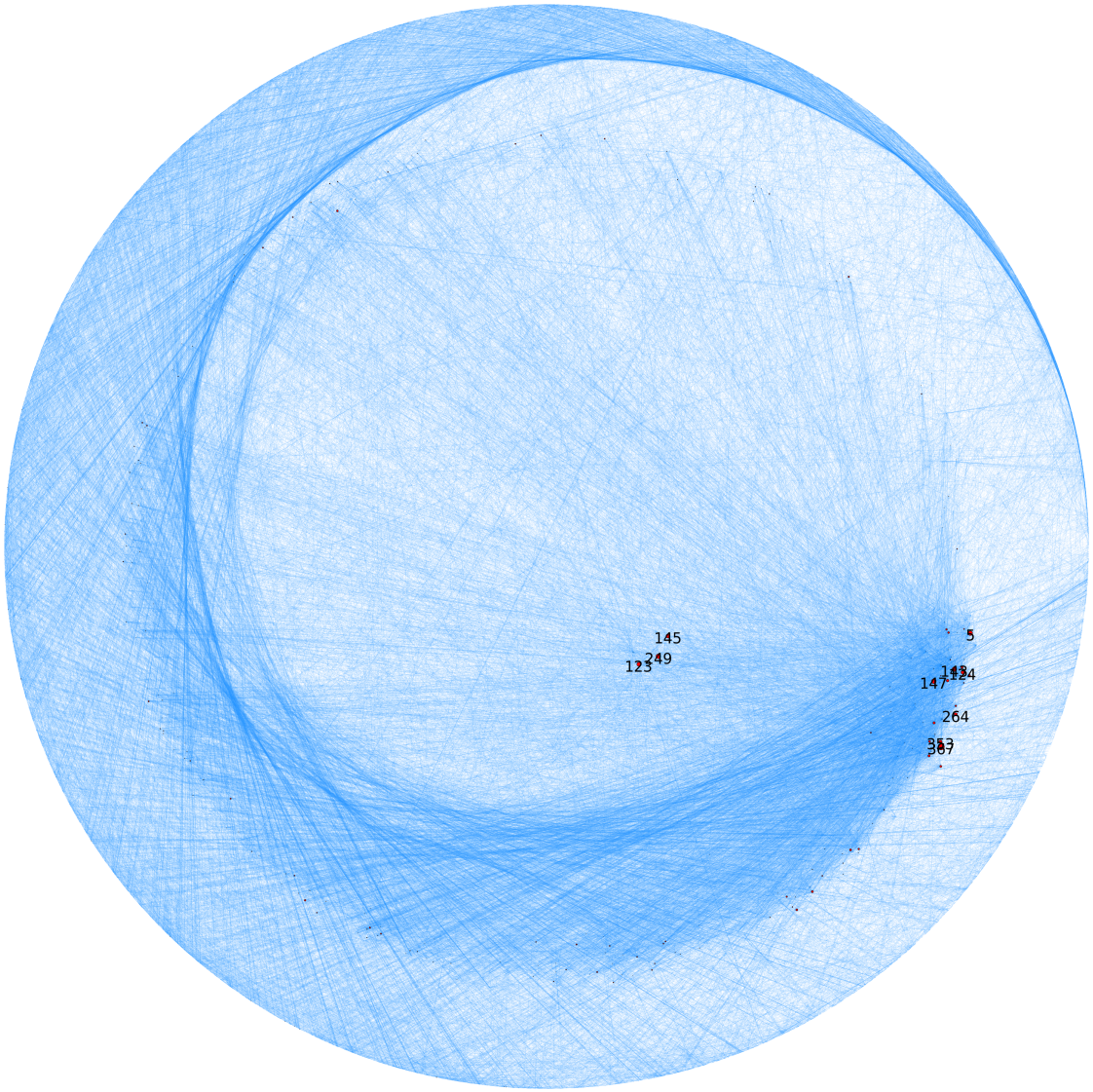
Facebook: Some facebook network plot with top 10 central nodes of information flow. [The numbers correspond to Node IDs.]



Size of the node represents the importance of information flow through the node.(Radius of nodes is proportional to the reconstructed egocentric local betweenness).

GNUTELLA: Some p2p Gnutella network plot with top 10 central nodes of information flow. [The numbers correspond to Node IDs.]

Interesting inference: In general p2p networks used to hash and distribute the node IDs in a uniform distribution. While sharing files/information they used to search using connections with their connected peers. For some reason our plotting algorithm also seems to distribute the nodes in a uniform distribution and it is clearly visible which nodes are hotspots and contributing the most to information flow.



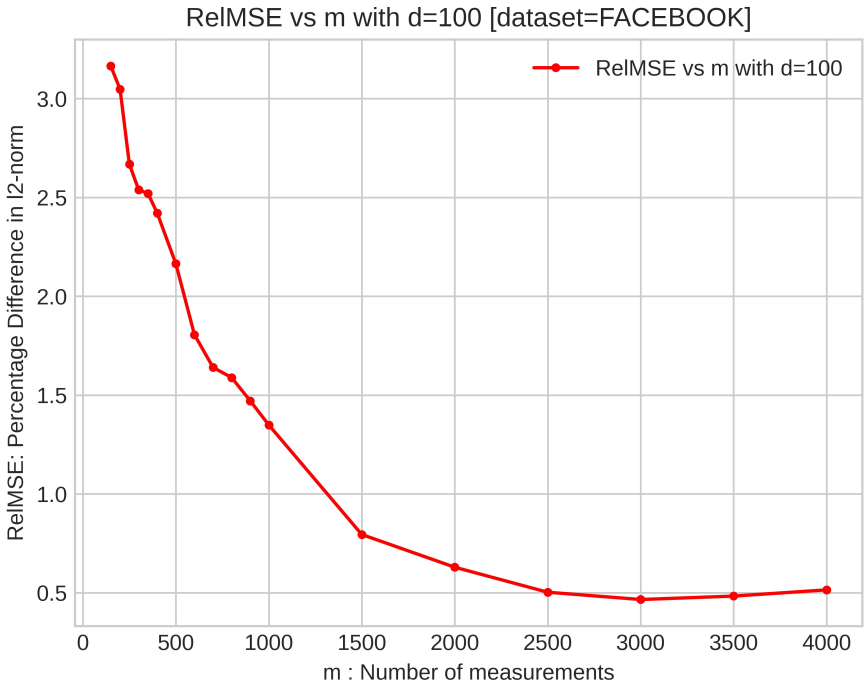
Size of the node represents the importance of information flow through the node.(Radius of nodes is proportional to the reconstructed egocentric local betweenness).

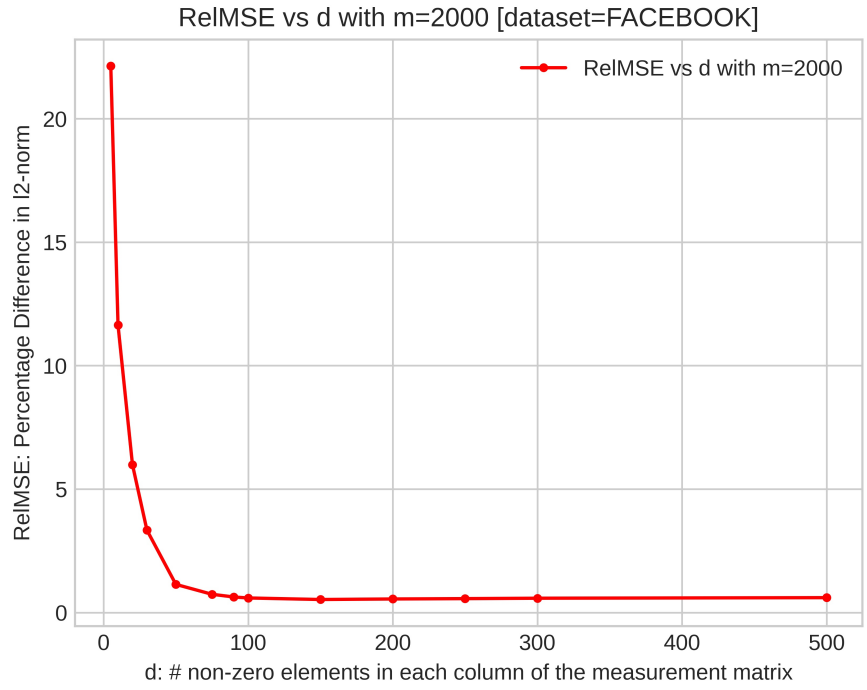
Error vs Measurement Parameters

We have measured two types of errors: **RelMSE**(Relative MSE: For measuring l2-norm based in reconstruction error) and **CatErr**(Categorization Error: Number of mismatches in identifying TopK nodes of information flow)

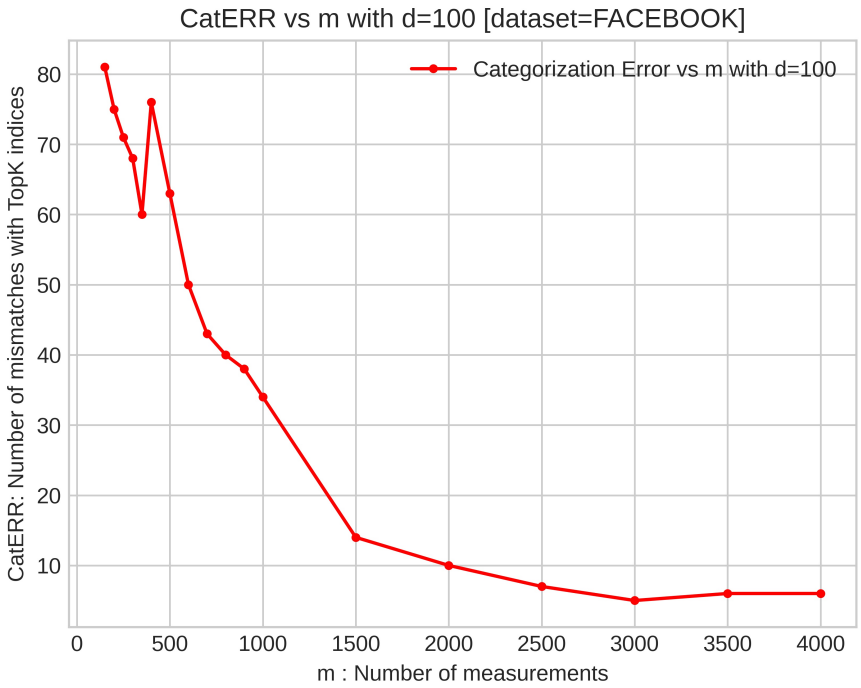
We run multiple tests on **Facebook** and **Gnutella** datasets with fixed **d** and fixed **m** separately and tabulate those results.

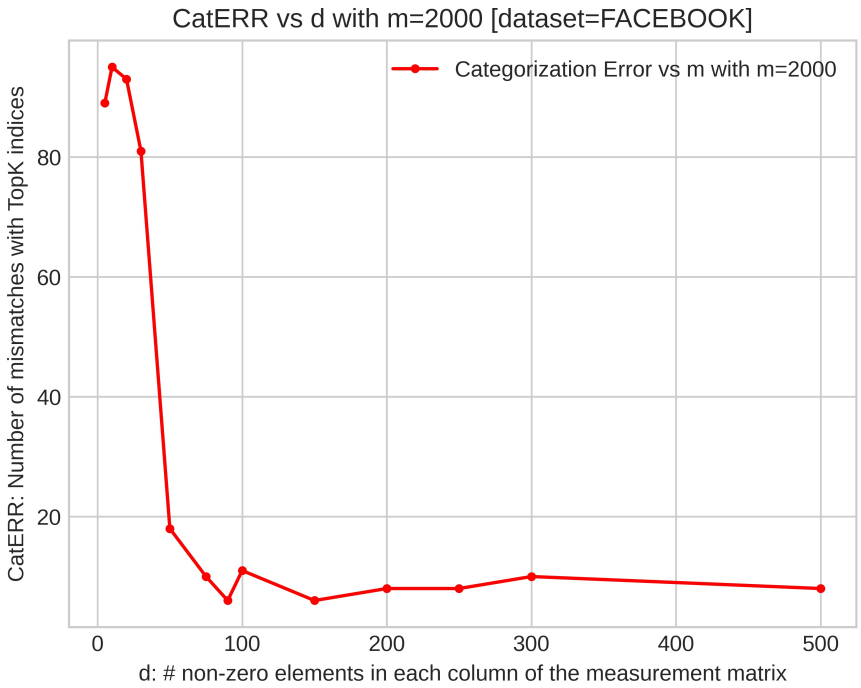
Facebook:





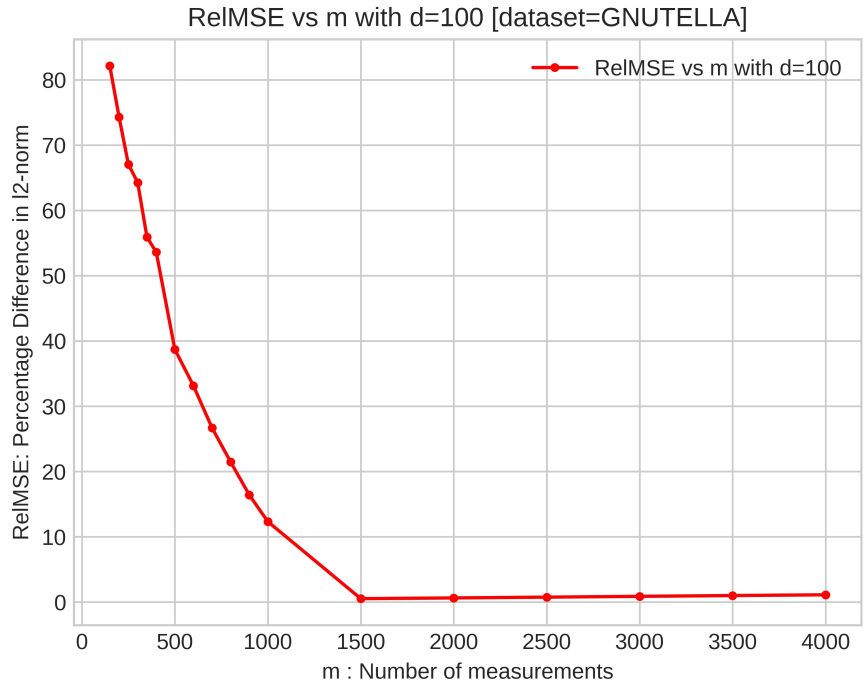
Categorization Error plots for Facebook dataset[**TopK=100**]:

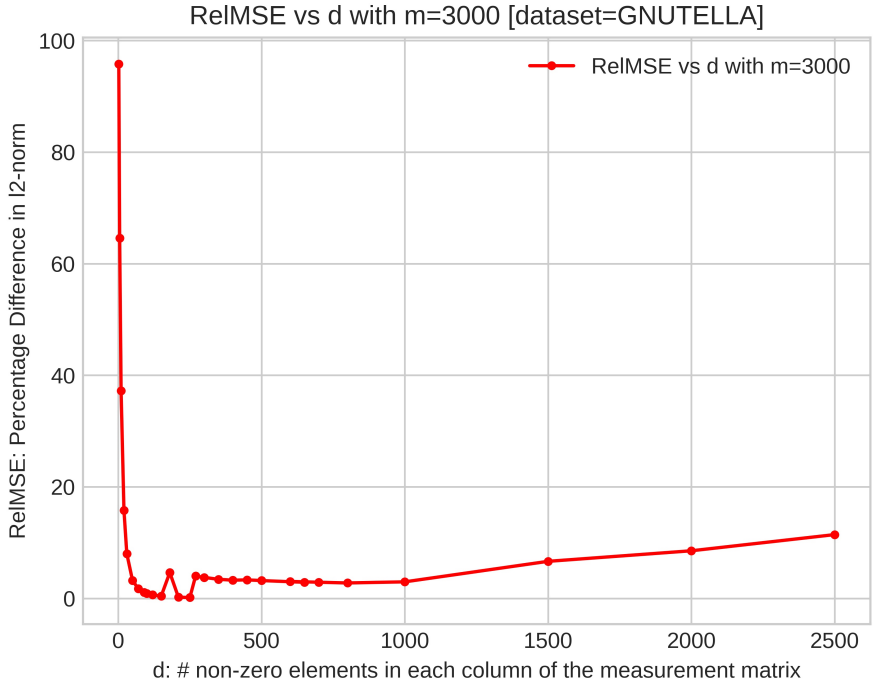




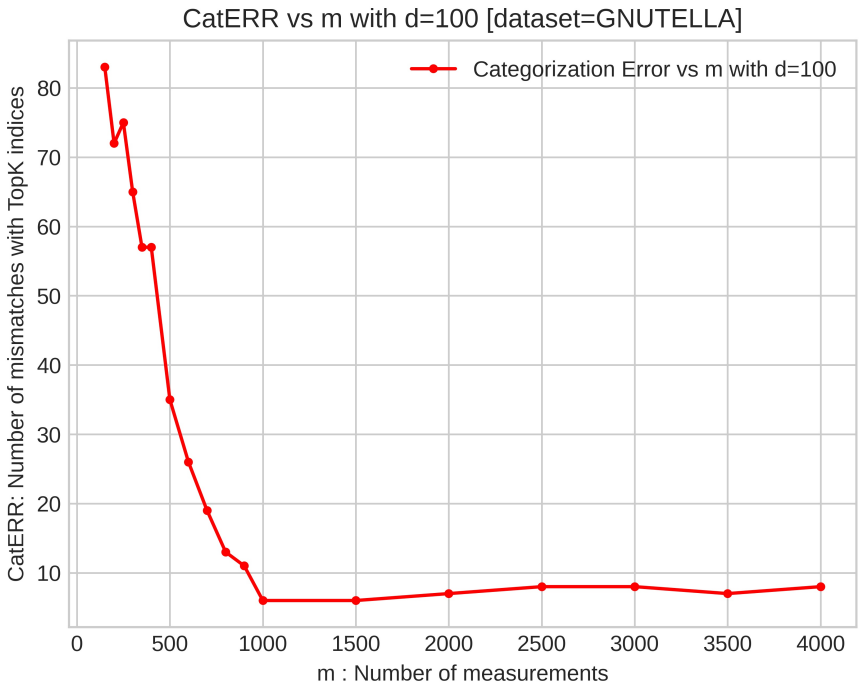
GNUTELLA:

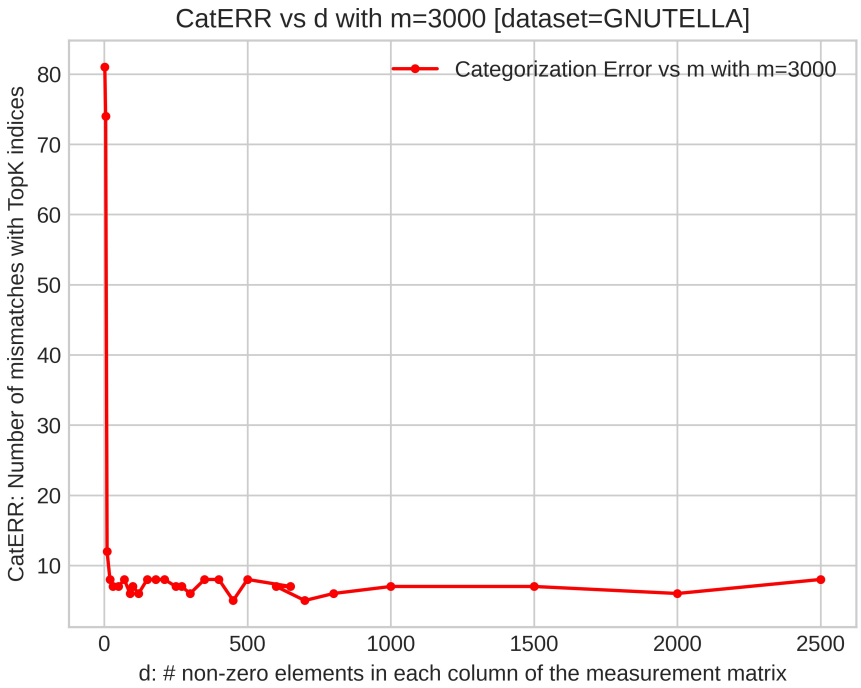
Relative MSE between the true centrality scores and the reconstructed scores.





Categorization Error plots for GNUTELLA dataset[**TopK=100**]:





Some More Applications of this type of Sensing Matrix:

Infection Propagation

Suppose that we have a large population where only a small number of people are infected by a certain viral sickness (e.g., a flu epidemic). The task is to identify the set of infected individuals by sending agents among them. Each agent contacts a pre-determined or randomly chosen set of people. Once an agent has made contact with an infected person, there is a chance that he gets infected, too. By the end of the testing procedure, all agents are gathered and tested for the disease. It is realistic to assume that, once an agent has contacted a person, the next contact will be with someone in close proximity of that person. Therefore, in this model we are given a random geometric graph that indicates which set of contacts can be made by an agent.Now, the question here is to determine the number of agents that is needed in order to identify the set of infected people.

Sensor Networks

A sensor network is static with a given graph topology such as a geometric random graph. Sensor networks can be monitored passively or actively. In passive monitoring, at any instant, sensor nodes form a tree to route packets to the sink. The routing tree constantly changes unpredictably but must be consistent with the underlying network connectivity. A test is considered positive if the arrival time is significantly large, which indicates that there is at least one defective sensor node or a congested link. The goal is to identify defective links or sensor nodes based on packet arrival times at the sink. In active monitoring network nodes continuously calculate some high level, summarized information such as the average or maximum energy level among all nodes in the network. When the high level information indicates congested links, a low level and more energy consuming procedure is used to accurately locate the trouble spots