# Graph Compression: One-Year Report

Suvidha Kancharla, Jen Neville, Thomas Courtade, and Victoria Kostina

## I. Problem Description

Estimating models of graph data from network samples is a task of fundamental importance for many problems today. For example, computing page rank for search engines, or peer influence in social networks both rely on an estimate of the underlying network model. However, in many cases the underlying graph is unknown, and must be estimated through sampling. One of the most fundamental sampling mechanisms are so-called "random walks", which are broadly applied in graph sampling with applications to web search [3], community identification [5], and PageRank computation [4].

In this project, we sampled graphs (both real and synthetically generated) via random walks, and studied the effect of sampling on the compression rate required to store the graph in memory. On that note, many present graph compression algorithms are heuristic in nature and are aimed toward

1) Reducing the storage space by designing succinct data structures
2) Reducing the time taken for accessing user queries by storing graphs in data structures which facilitate fast random access to an edge.

In this project, we approached the first point from a different perspective. That is, we studied the performance of a popular graph compression algorithm, LLP [2], against the performance of the algorithm presented in [1] on real and synthetic datasets. This concretely demonstrated that current algorithms are far from theoretically optimal, and hence there is significant potential to improve current compression technology for social networks.

## II. Progress and insights thus far

The paper [1] proves bounds on structures induced by the Erdös-Rényi random graph model. Suvidha has performed many experiments which check the compression of the algorithm posed in [1] on social graphs. The compression rate was then empirically compared against the popular Layered Label Propagation (LLP) framework [2].

In particular, we perform a random walk for a fixed number of steps and the obtained subgraphs are passed to the compression algorithms of [1] and LLP. The random walks do not necessarily obtain the full graph, so we use this as a means to explore how lossy compression behaves.

For a variety of social graphs, we have computed the empirical rate-distortion performance (see appendix for experimental results). Distortion is taken to be the number of edges lost (i.e., the Jaccard coefficient). The number of bytes needed to compress the distorted graph is used as a proxy for compression rate.

Although the algorithm in [1] is proved to be optimal for the Erdös-Rényi random graph model, it also outperforms LLP on the real social graphs examined. Since the algorithm in [1] is not expected to exploit power-law structure to its fullest extent, the compression rate it attains is an upper bound on that which is theoretically optimal. Thus, there is much work to do – and much to be gained – in terms of designing clever compression algorithms for real networks.

The results are most succinctly illustrated by the scatter plot in Figure 1. In Figure 1, the blue dots represent the normalized compression ratios attained by LLP for the labeled datasets, and the aligned red dots are the corresponding normalized compression ratios attained by the graph compression algorithm which was inspired by [1]. The normalized compression ratio is computed

by dividing the number of bits it requires to represent the compressed graph by the number of bits required to represent a memoryless sequence of bits with probability of one equal to the marginal edge probability in the given graph (i.e., $H(p) \times n(n-1)/2$).

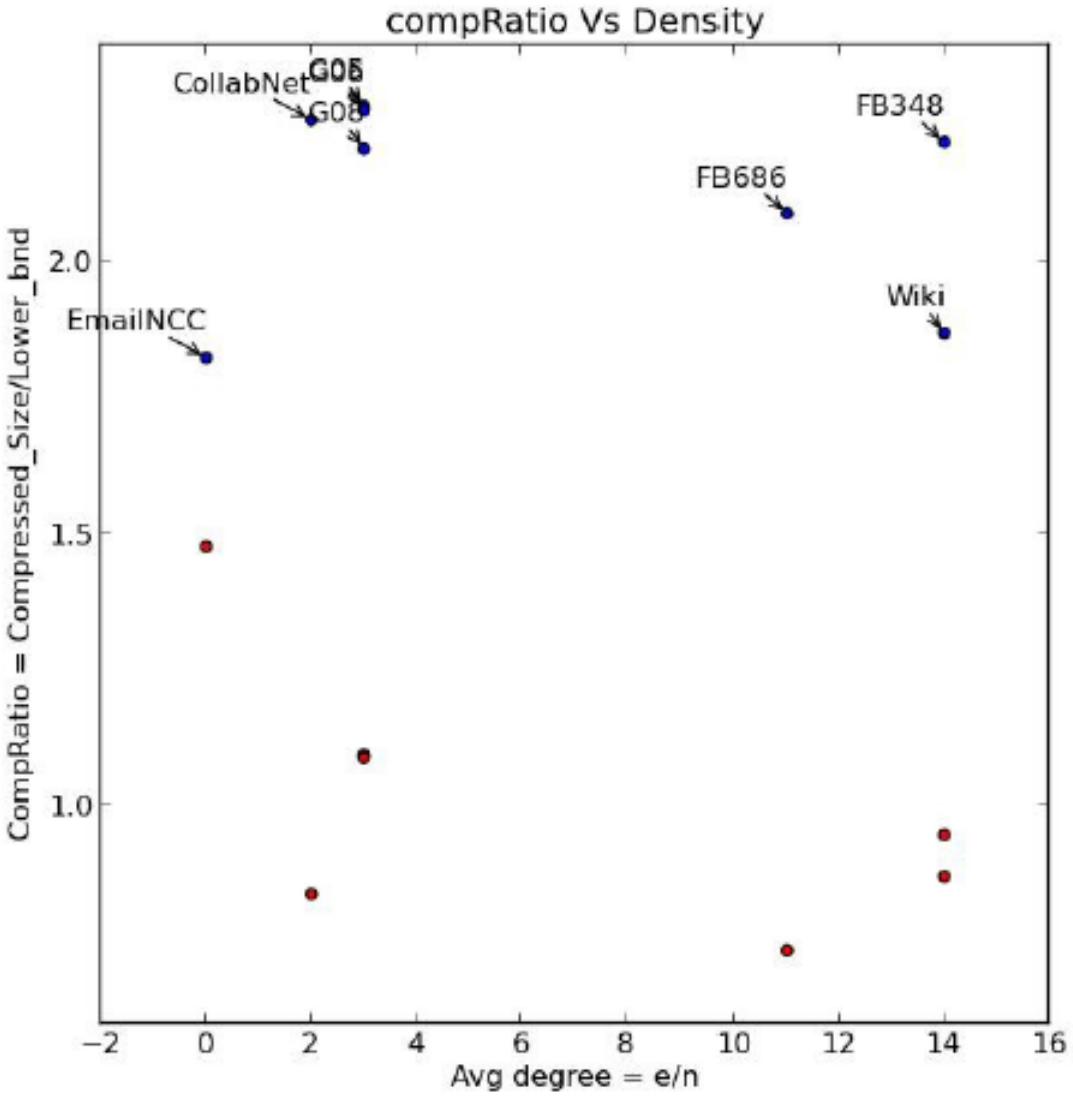



Fig. 1. A comparison of compression ratios attained by LLP and the algorithm described in [1] for real social networks.

The datasets included in Figure 1 are as follows:

- CollabNet Graph: http://snap.stanford.edu/data/ca-GrQc.html
- Facebook Graph (FB348, FB686): http://snap.stanford.edu/data/egonets-Facebook.html
- Purdue Email Graph
- Gnutella peer to peer network (G05,G06,G08): http://snap.stanford.edu/data/p2p-Gnutella0X.html
- Wikipedia Voting dataset: http://snap.stanford.edu/data/wiki-Vote.html

Complete results can be found in the attached notes which documents the experiments which were performed. In particular, the attached experimental results demonstrate that the algorithm inspired by [1] continues to outperform LLP in the lossy regime (i.e., when the entire graph is not sampled by a random walk). Figure 2 supplies an example corroborating this claim for the Purdue Email Graph. In general, we have observed that the trends are robust and do not significantly vary between datasets.

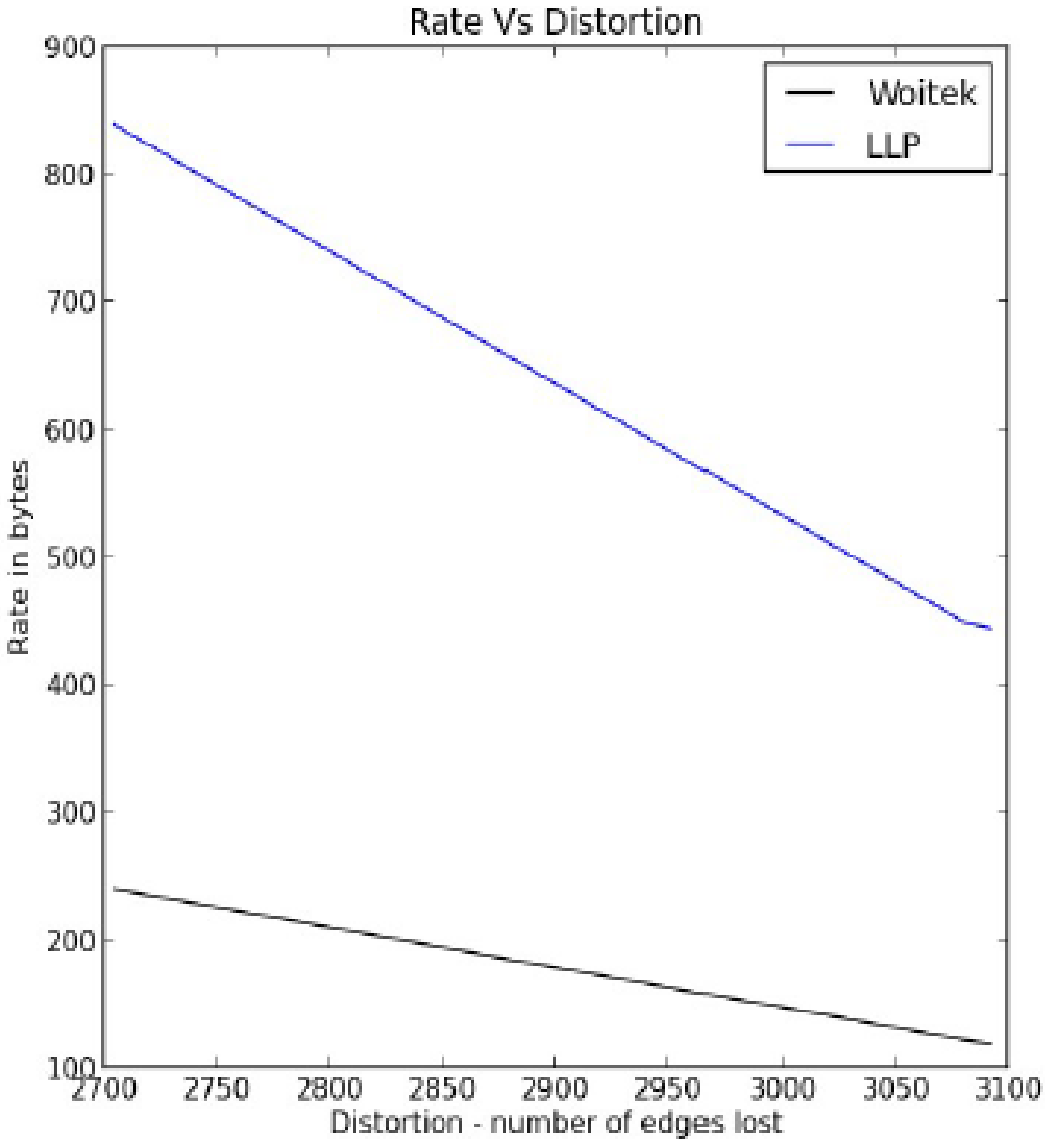



Fig. 2. A comparison of compression ratios attained by LLP and the algorithm described in [1] for lossy versions of the Purdue Email Graph obtained via 10 independent, parallel random walks.

## III. Outcomes: Poster Presentations

• Suvidha Kancharla, "*Graph Inference Based on Random Walks*," Poster presentation at the December, 2012 NSF site visit at Purdue University.

## IV. Team members' experience regarding interdisciplinary interactions

This project brought together researchers from theoretical and empirical backgrounds. Notably, Tom's background in information-theoretic compression and Jen's experience using real datasets combined nicely to guide a sequence of experiments on real data which compared the performances of a theoretically optimal compression algorithm [1] against that of a popular, heuristically-designed algorithm [2]. As a result, we discovered that the popular LLP algorithm [2] used for social and web graph compression performs significantly worse than the optimal compression algorithm [1] does. Moreover, this trend holds even when the graphs are sample in a lossy manner. This demonstrates that there is significant work to be done in this area.

## V. Team meetings

- Period of September, 2012 - March 1, 2013.
  - Suvidha and Jen meet regularly in person to discuss this project.
  - Suvidha, Jen, and Tom met at the NSF site visit in December. At this workshop, research directions were discussed and Suvidha presented a poster.
  - Tom and Victoria meet regularly, but have not been discussing work immediately related to the work discussed in this report. Instead, their discussions have focused on the finite-blocklength compression aspects of the initial proposal, without the graph component (since finite-blocklength itself is extremely challenging).
- Period of March, 1 2013 - present.
  - Skype Telecons were held approximately once per month between Tom, Jen, and Suvidha to discuss progress.
  - Tom and Jen held an in-person meeting at the Big Data workshop in Hawaii in mid-March.
  - Tom and Jen held an in-person meeting at Purdue in June at the IT Summer School.
  - Tom and Jen will meet regularly at the Big Data Program held at Berkeley's Simons Institute in Fall, 2013. They plan to discuss how the current work can be leveraged in order to produce a conference paper.

## VI. Budget and Plans to Continue

Of the initial $5K budget, only $721.85 was spent (to support Tom's travel to the Information Theory Summer School at Purdue, where Tom & Jen held an in-person team meeting). Given that we were not able to justify using the entire budget (see also Section VII below), we do not plan to to request funds for an additional year. However, if Tom and Jen are successful in turning this work into a conference paper this fall, some of the remaining budget may be used to attend the conference. In any event, the team thanks CSoI for providing us with this opportunity to interact and share ideas.

## VII. Comments for improving future teams

In order for me to comment on improving the seed grant experience for future teams, let me provide some context:

The greatest hurdle we faced on this project was student participation. In particular, despite Victoria's interest in the project, it became clear from early on that she did not have the bandwidth to participate. Indeed, she was in the final year of her Ph.D. and was fully occupied with writing her dissertation and wrapping up ongoing work with her advisor, both of which had higher priority. Thus, much of the work fell to Suvidha (the only remaining student) with Jen and Tom acting in advisory roles and guiding her work/experiments. Not only did this lead to progress which was slower than expected, but it also failed to promote student interaction, which I viewed as one of the chief goals of the seed grant projects.

I think the above situation highlights one thing that could be improved for next year. In particular, the advisor of each student involved in a seed grant should be listed as an advisor on the seed grant project and should play an active role in the proposal. This will ensure some commitment on the advisor's part to devote their student's time to the seed grant project. Moreover, it will make sure that the seed grant project a student is involved in does not conflict with their own advisor's plans for them. In short, the seed grant program is an excellent idea, and it provides a good platform for students to lead their own projects. However, it could benefit from including the involved students' advisors so that expectations are managed and do not conflict.

## REFERENCES


[1] Yangwook Choi and Wojciech Szpankowski, *Compression of Graphical Structures: Fundamental Limits, Algorithms, and Experiments*, 2011.

[2] Paolo Boldi, Marco Rosa, Massimo Santini, and Sebastiano Vigna. *Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks.* In Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, editors, Proceedings of the 20th international conference on World Wide Web, pages 587-596. ACM, 2011.

[3] Sergey Brin, Lawrence Page, *The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems*, Volume 30, Issues 17, April 1998, Pages 107-117.

[4] Nick Craswell and Martin Szummer. *Random walks on the click graph*. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2007). ACM, New York, NY, USA, 239-246.

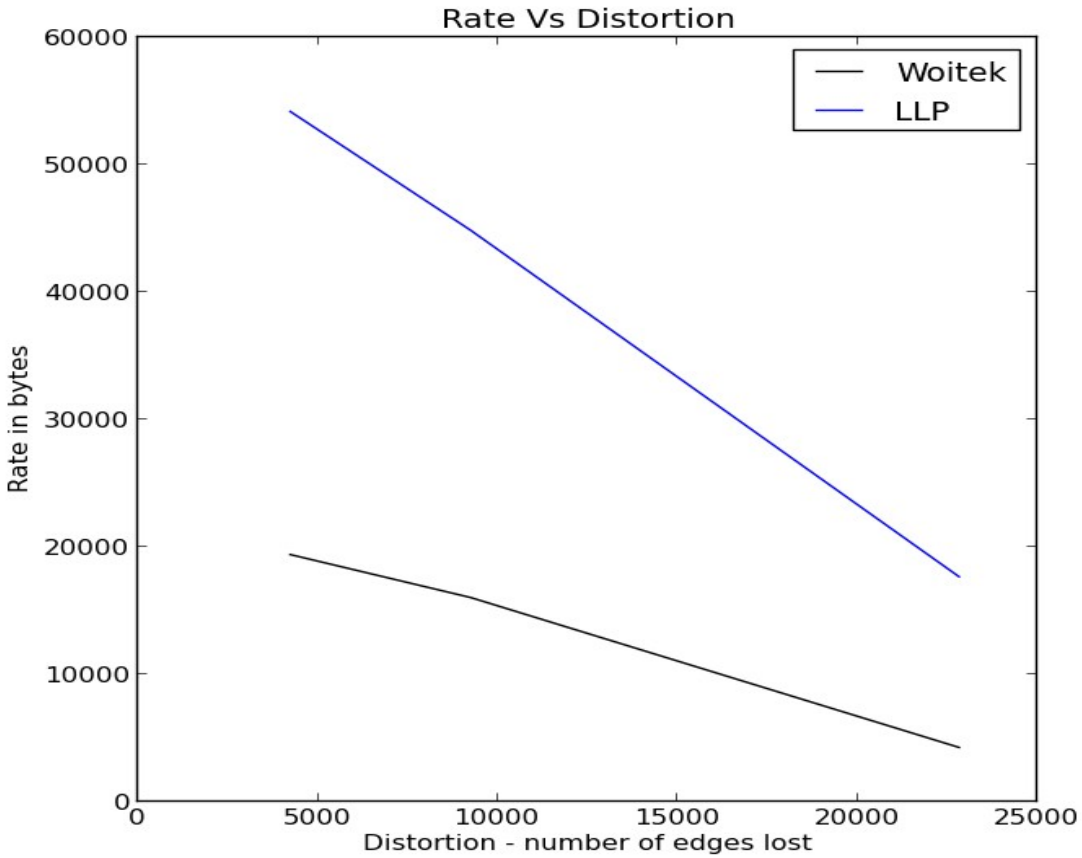
[5] Jiayuan Huang, Tingshao Zhu, and Dale Schuurmans. *Web communities identification from random walks*. In Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases (PKDD 2006), Johannes Frnkranz, Tobias Scheffer, and Myra Spiliopoulou (Eds.). Springer-Verlag, Berlin, Heidelberg, 187-198.

# Appendix: Experimental Results

Results so far:

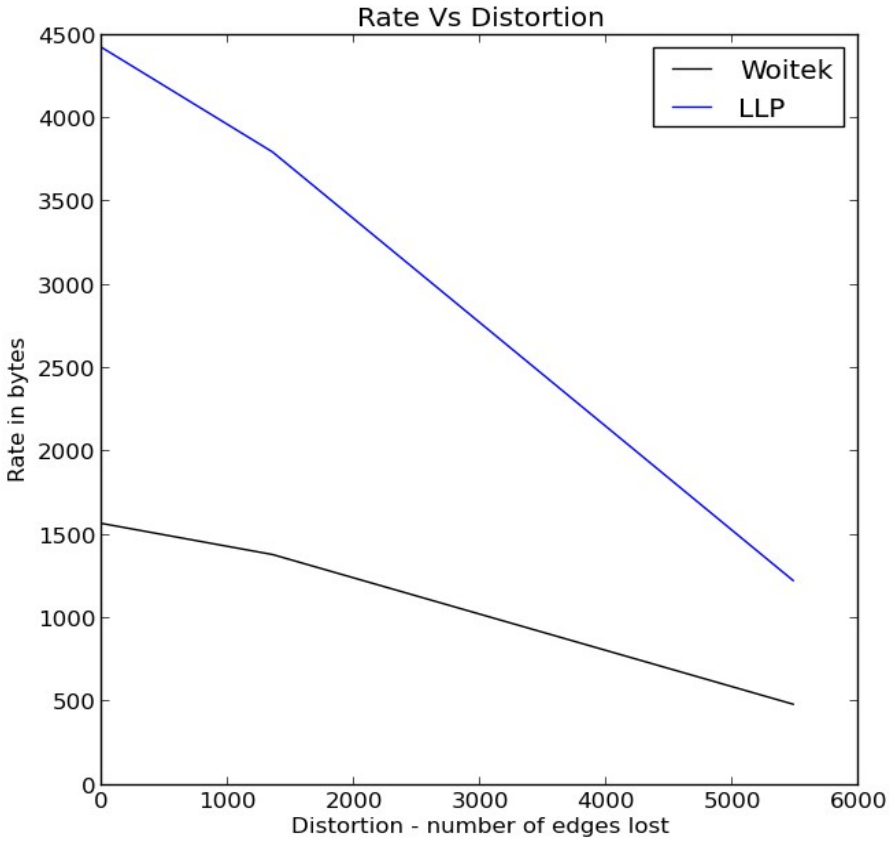
CollabNet Graph: (http://snap.stanford.edu/data/ca-GrQc.html)

| |
|---|
| TotalNodes - 5242 |
| Total Edges - 28980 |
| EdgesTravelled in [1000, 5000, 10000]steps by 10 random walkers : [6166, 19733, 24753] |
| Distortion in bytes [22814, 9247, 4227] |
| Wojciech rate: [4237, 15998, 19375] |
| LLP  rate: [17636, 44814, 54148] |



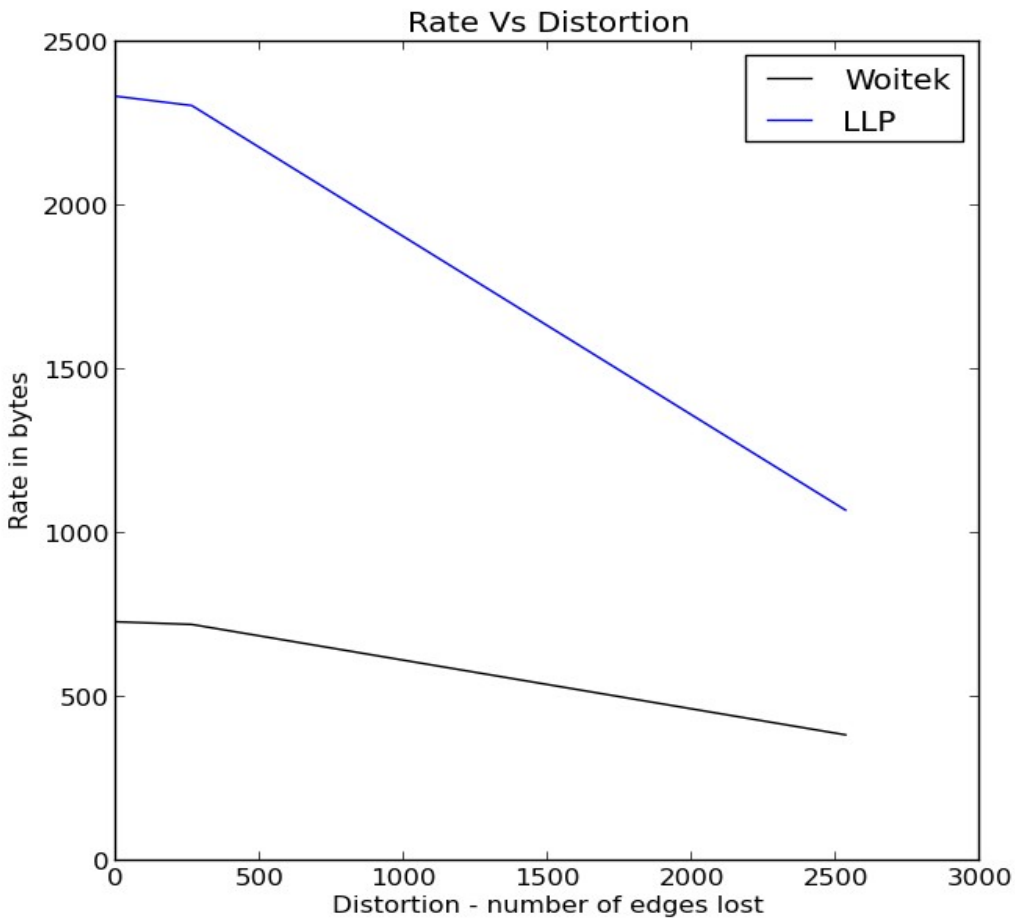
2)FB348 (http://snap.stanford.edu/data/egonets-Facebook.html)

| |
|---|
| Total Edges : 6384 |
| Total Nodes : 224 |
| EdgesTravelled in [100, 1000, 5000]steps by 10 random walkers :  [906, 5033, 6383] |
| Distortion in bytes: [5478, 1351, 1] |
| Wojciech rate in bytes: [486, 1384, 1570] |
| LLP in bytes:[1228, 3799, 4424] |



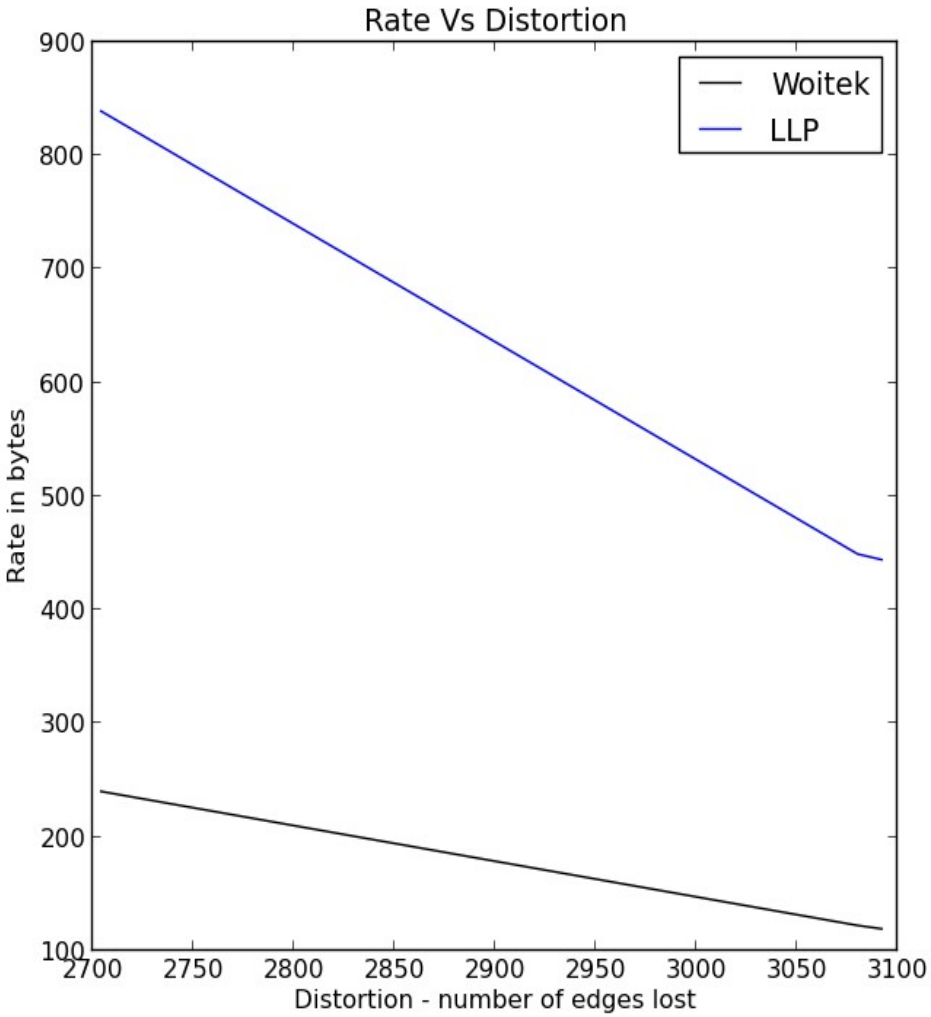
3)FB686 ((http://snap.stanford.edu/data/egonets-Facebook.html))

| |
|---|
| Total Edges 3386 |
| Total Nodes 150 |
| EdgesTravelled by 10 random walkers in [100, 1000, 5000] steps : [854, 3124, 3384] |
| Distortion [2532, 262, 2] |
| Wojciech : [384, 721, 729] |
| LLP : [1070, 2305, 2333] |

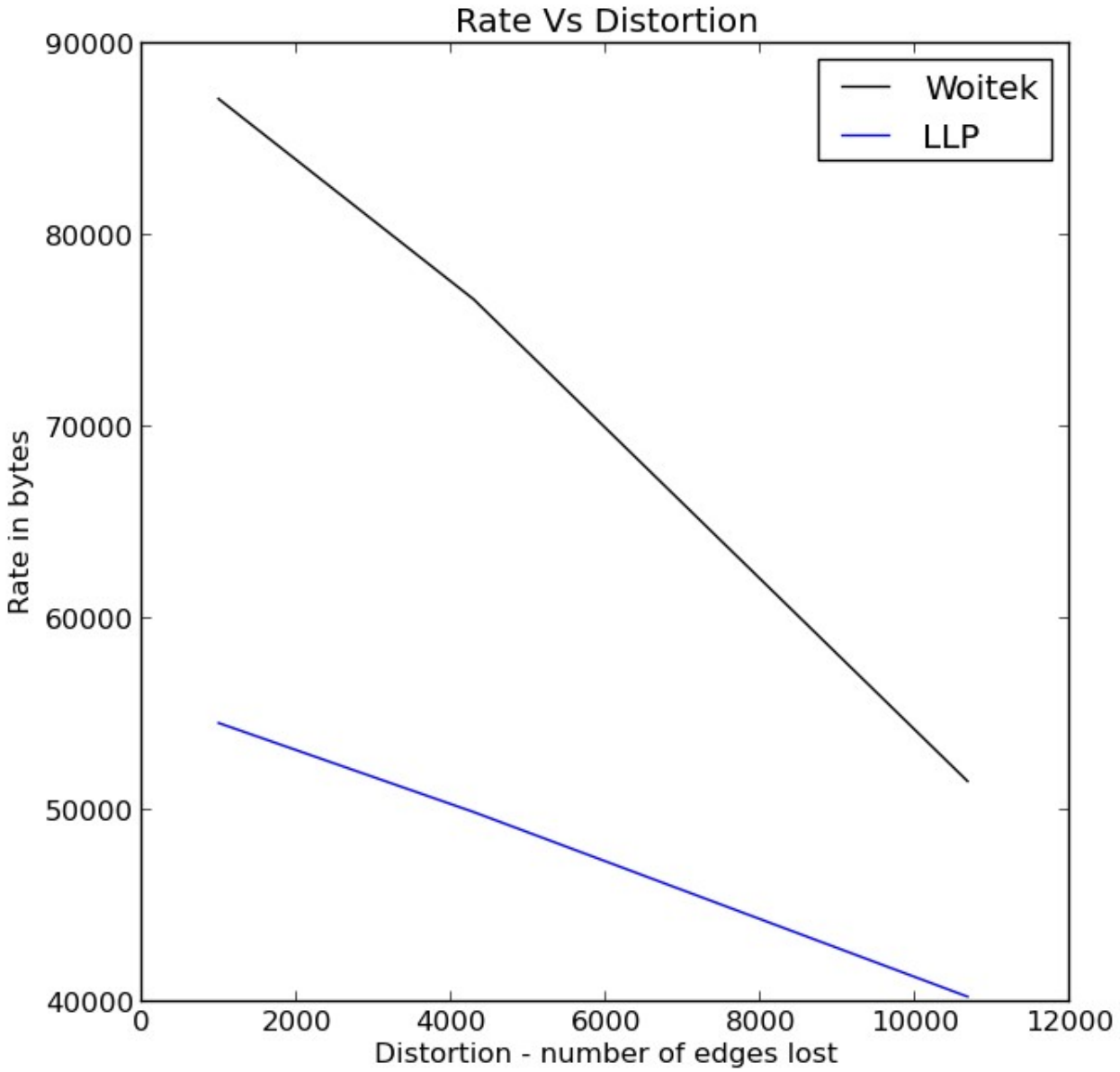


4) Purdue Email Graph

| |
|---|
| Total Edges 3228 |
| Total Nodes 2187 |
| EdgesTravelled in [100, 500, 1000]steps by 10 random walkers [136, 148, 524] |
| Distortion : [3092, 3080, 2704] |
| Wojciech : [119, 122, 240] |
| LLP : [444, 449, 839] |



p2p-Gnutella06.txt
Total Edges31525
Total Nodes8717
EdgesTrav in [10000, 20000, 40000]steps for10 random walks:[20855, 27243, 30542]
Distortion[10670, 4282, 983]
Woitec : [51505, 76651, 87111]
LLP :[40263, 49897, 54543]



emailNCC
Total Edges3228
Total Nodes2187
Woitech
[1129]
LLP
[6417]

---

0.0366420274675

---

Fb686
Total Edges3386
Total Nodes150
Woitech
[1108]
LLP
[2339]

---

0.670292023834

---

FB348edges
Total Edges6384
Total Nodes224
Woitech
[2477]
LLP
[4424]

---

0.54428147097

---

CollabNet
LLPGraph63400
Total Edges28980
Total Nodes5242
Woitech
[43881]
LLP
[63400]

---

0.529635811052

---

Gnutella-08.txt
Total Edges20777
Total Nodes6301
Woitech
[43378]
LLP
[35009]

---

0.0108679219358

---

Gnutella-05
LLPGraph56901
Total Edges31839

Total Nodes8846
Woitech
[88278]
LLP
[56901]
0.0072010656711

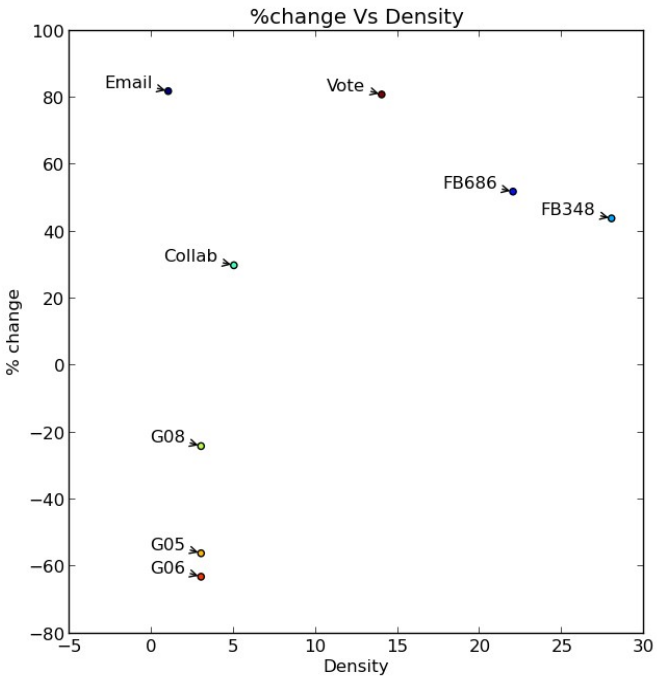
---

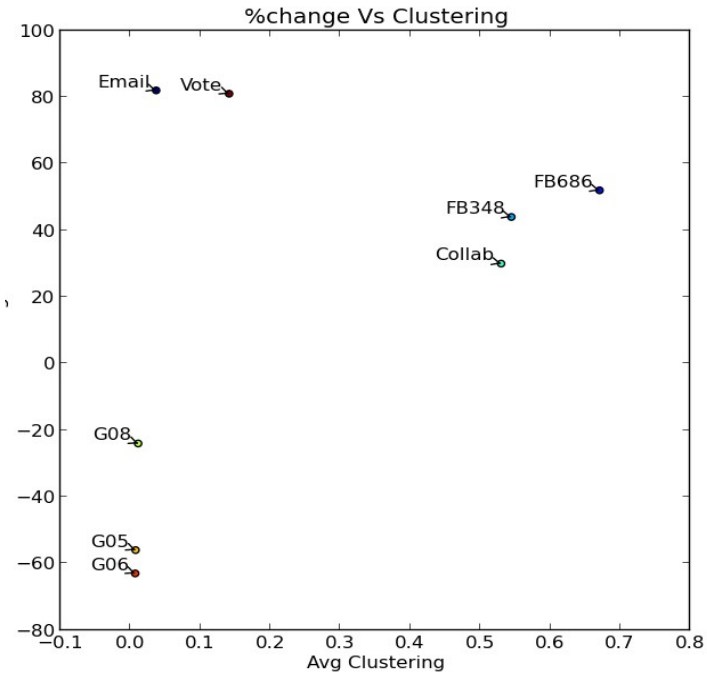
Gnutella06
LLPGraph55894
Total Edges31525
Total Nodes8717
Woitech
[91012]
LLP
[55894]
0.00667646455292

---

Wiki-Vote
Total Edges103689
Total Nodes7115
Woitech
[24634]
LLP
[132897]
0.140897845893

---


%change Vs Density
Email
Vote
FB686
FB348
Collab
G08
G05
G06
% change
Density
100
80
60
40
20
0
−20
−40
−60
−80
−5
0
5
10
15
20
25
30


%change Vs Clustering
Email
Vote
FB686
FB348
Collab
G08
G05
G06
Avg Clustering
100
80
60
40
20
0
−20
−40
−60
−80
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

ErdosRenyi:
4000 nodes , p =0.15
Total Edges1198433
Total Nodes4000
Woitech
[877194]
LLP
[868168]
0.14989448915

---

ErdoesRenyi
p=0.001
clustering:0
Total Edges3
Total Nodes100
Woitech
[100]
LLP
[18]

Total Edges46
p=0.001
clustering : 0
Total Nodes100
Woitech
[122]
LLP
[80]

p=0.05
Total Edges284
Total Nodes100
clustering: 0.0670158730159
Woitech
[433]
LLP
[315]

p=0.075
Total Edges350
Total Nodes100
clustering:0.0835137640138
Woitech
[464]
LLP
[383]

p=0.1
clustering : 0.103325703708
Total Edges500
Total Nodes100
Woitech
[555]
LLP
[487]

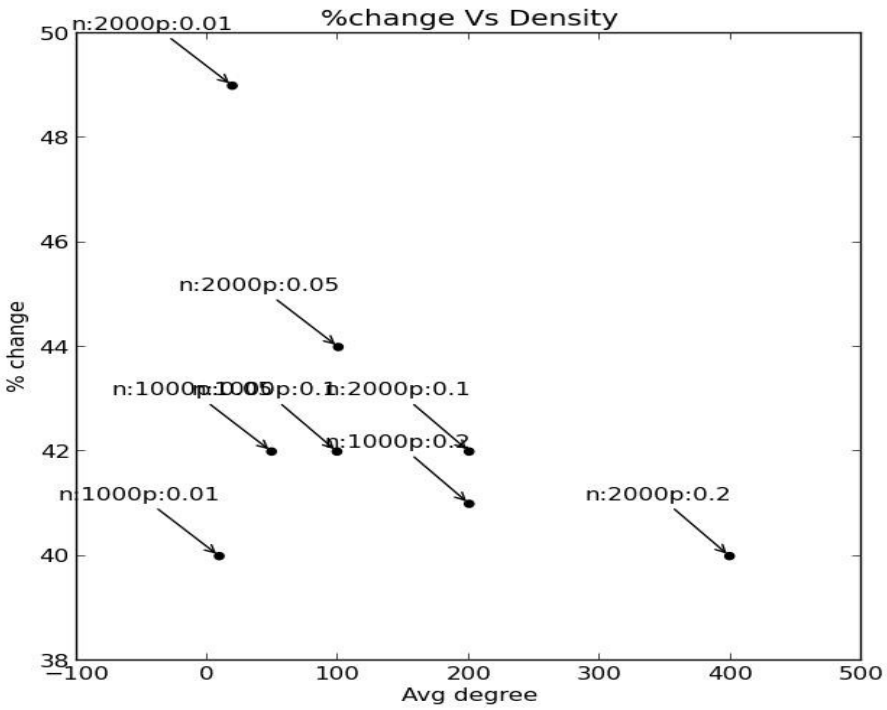
p=0.3
clustering:0.304223468584
Total Edges1464
Total Nodes100
Woitech
[1000]

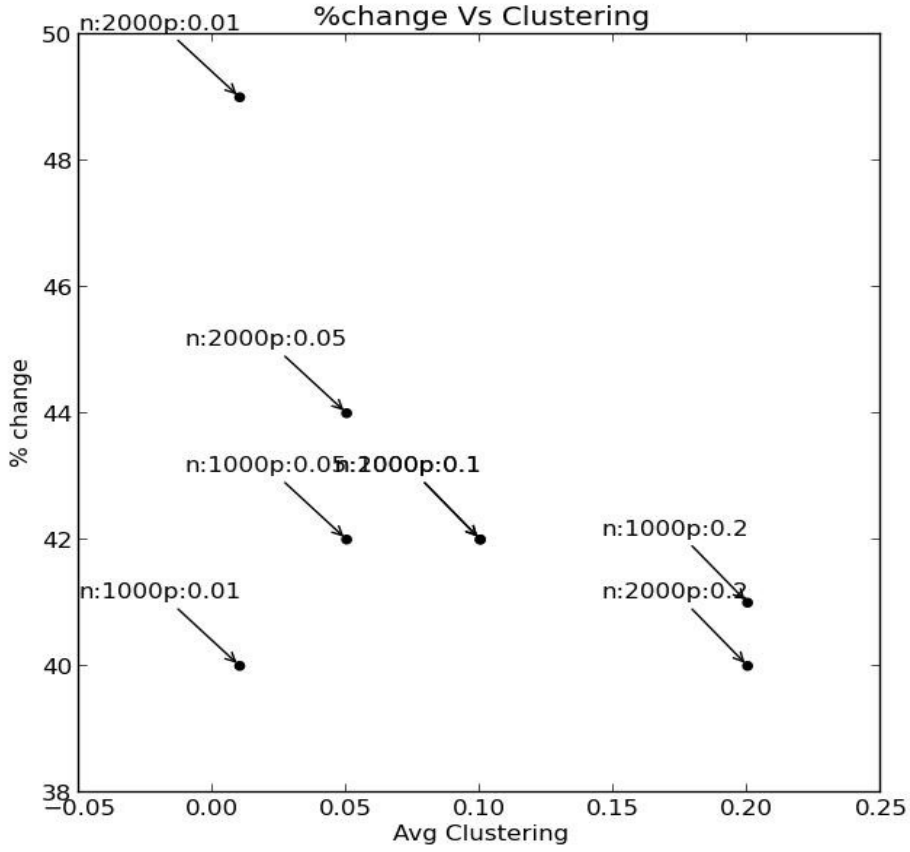
LLP
[1007]

100
2411
p=0.5
0.484781020261
Total Edges2411
Total Nodes100
Woitech
[1304]
LLP
[1349]

---

19th March -
Mistake identified : Erods Renyi plots were previously undirected and LLP was directed. Now both compressions are directed. Woi is better over LLP for Erdos Renyi plots.

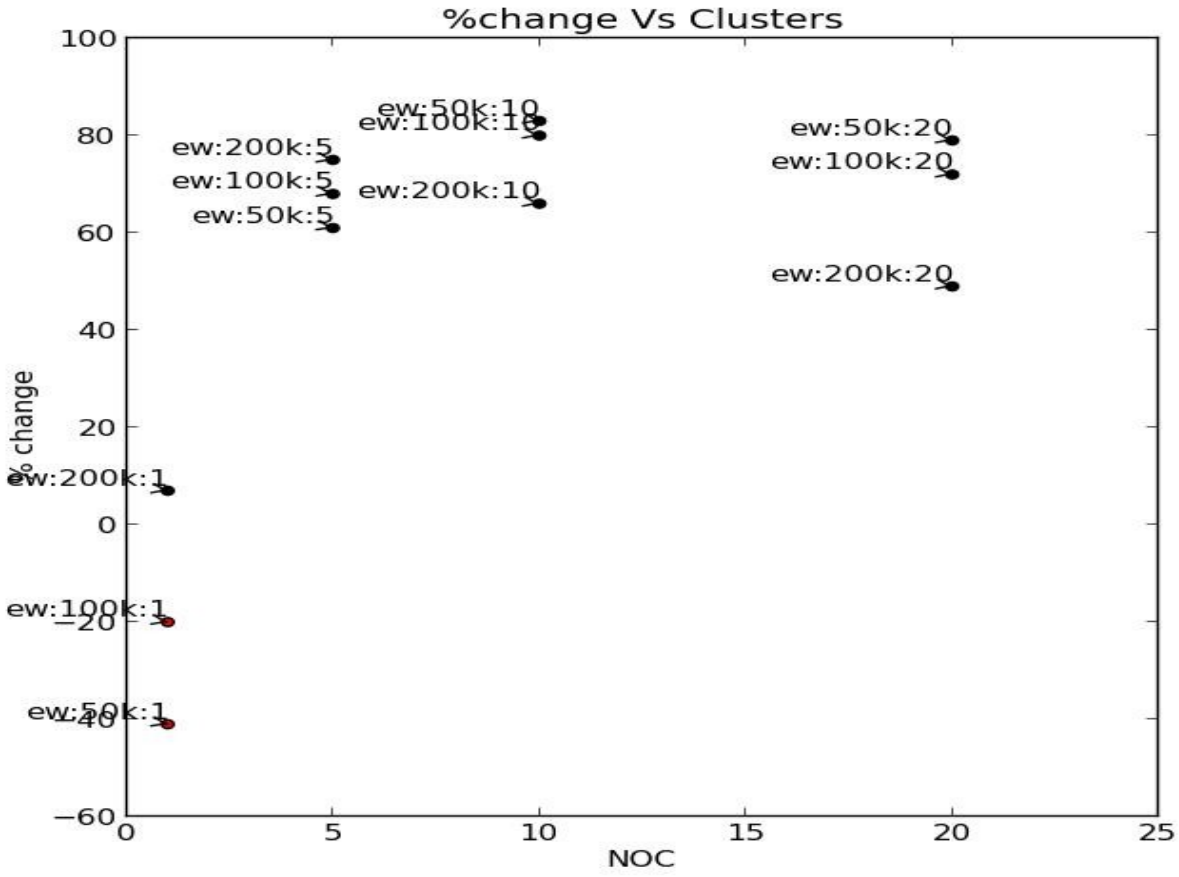

%change Vs Density
n:2000p:0.01
n:2000p:0.05
n:1000p:0.05
n:1000p:0.1
n:2000p:0.1
n:1000p:0.2
n:1000p:0.01
n:2000p:0.2
% change
Avg degree
50
48
46
44
42
40
38
−100
0
100
200
300
400
500

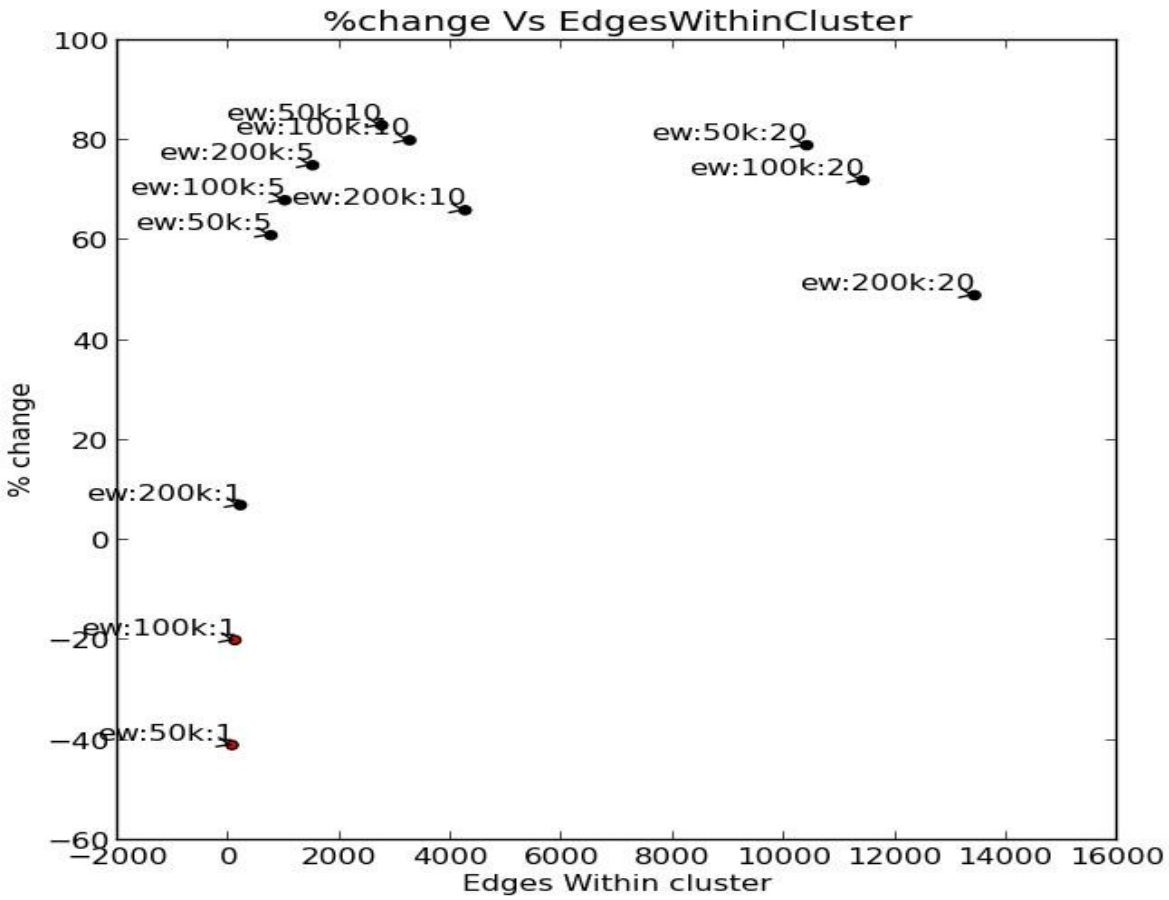

%change Vs Clustering
n:2000p:0.01
n:2000p:0.05
n:1000p:0.05
n:2000p:0.1
n:1000p:0.2
n:1000p:0.01
n:2000p:0.2
% change
Avg Clustering
50
48
46
44
42
40
38
−0.05
0.00
0.05
0.10
0.15
0.20
0.25

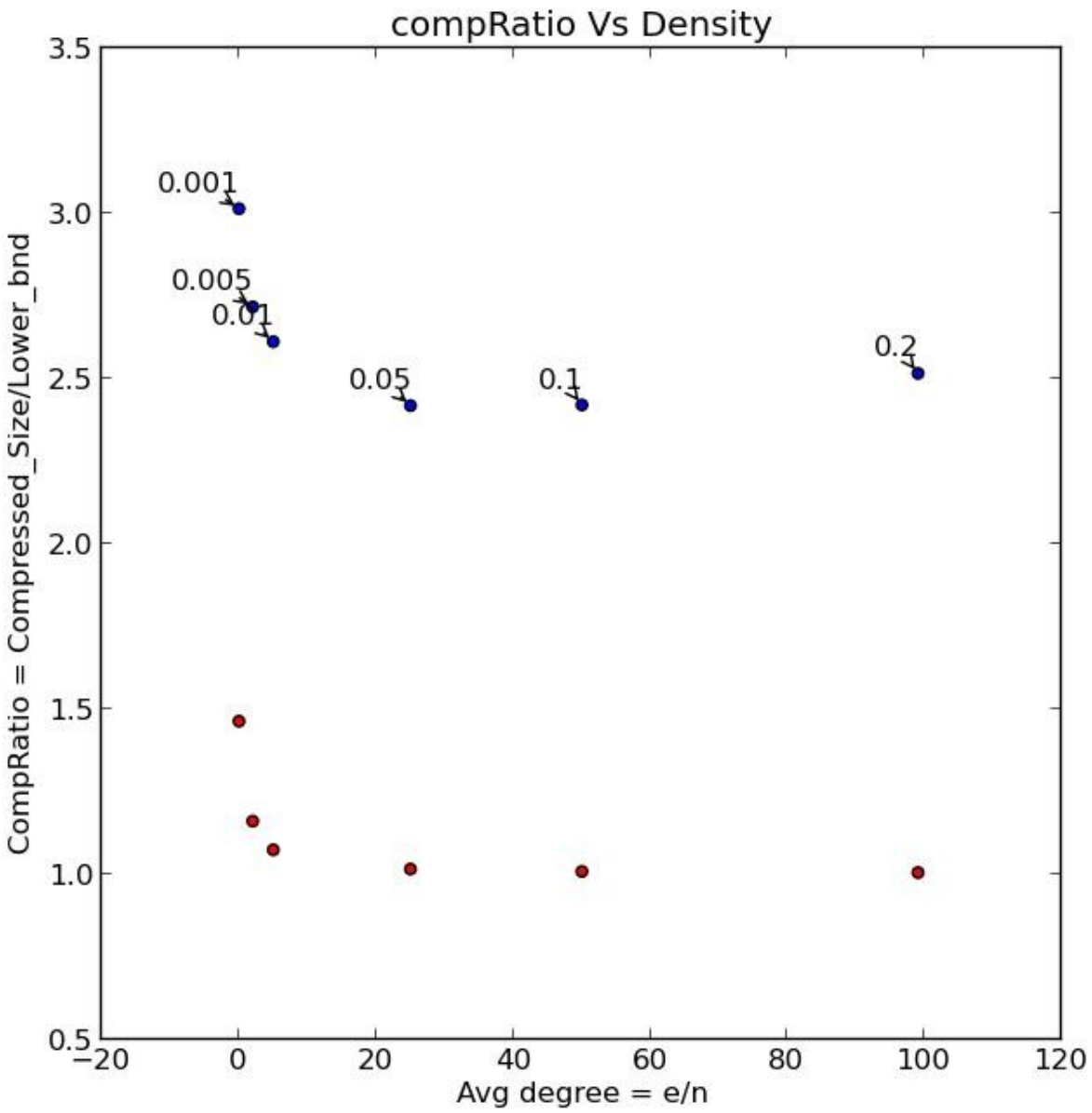
Custom

plots:

N: num of nodes = 1000
eb = edges between clusters = 10
ew = edges within cluster = [50,100,200]
nc= number of clusters = [1,5,10,20]




%change Vs EdgesWithinCluster
% change
Edges Within cluster
ew:50k:10
ew:100k:10
ew:200k:5
ew:100k:5
ew:200k:10
ew:50k:5
ew:50k:20
ew:100k:20
ew:200k:20
ew:200k:1
ew:100k:1
ew:50k:1
100
80
60
40
20
0
−20
−40
−60
−2000
0
2000
4000
6000
8000
10000
12000
14000
16000




Red: Woitech, Blue: LLP
Social Graphs:
**N**=[2187, 5242, 224, 150, 6301, 8846, 8717, 7115]
**E** =[1614, 14496, 3192, 1693, 20777, 31839, 31525, 100762]
**LLP Compression rate in bits**
[35256, 371072, 30568, 14320, 520216, 851784, 839024, 1773072]
**Woitech Compression rate in bits**
[28548, 137538, 11994, 5035, 257521, 407460, 400593, 897700]
**Entropy**:
[19327, 164114, 13771, 6856, 235647, 372677, 368115, 948404]

**Woitech Comp ratio**
[1.477, 0.838, 0.87, 0.73, 1.09, 1.093, 1.088, 0.95]
**LLP Comp ratio**
[1.82, 2.26, 2.22, 2.08, 2.20, 2.28, 2.27, 1.86]