Influence Study on Hyper-graphs

Dimitrios Vogiatzis
Institute of Informatics and Telecommunications, NCSR “Demokritos”
Athens, Greece
and
The American College of Greece (Deree)
6 Gravias Street, GR-153 42, Aghia Paraskevi
Athens, Greece

Abstract

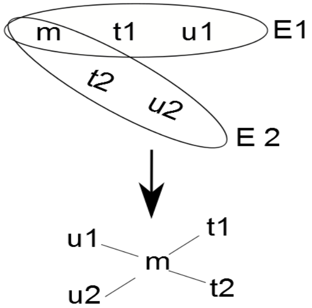
Multilateral relations between entities lose their semantics when represented as simple graphs. Instead hypergraphs can naturally represent the said relations, which are common in social tagging systems. An important issue is the effect of the structural properties of a hypergraph on influence propagation. In the current work, an empirical study is undertaken to compare the effect of degree, k-shell and eigenvector centrality under the SIS, and SIR models of infection. The results on the MovieLens, Delicious and LastFM social networks indicate that k-shell centrality is a more accurate predictor of the influence of a node than degree centrality, and that eigenvector centrality is closely correlated with k-shell centrality.

Introduction

A graph is perhaps the most popular representation of a social network. It represents dyadic relations between actors, where the edges stand for those relations and the nodes for the actors. However, there are many cases in which triadic or even multilateral relations need to be represented. For instance, a seller, a buyer and a broker participating in a business transaction; or a person seeing a movie, rating it, and annotating it with tags. Mapping triadic to dyadic relations leads to loss of information, for instance imagine two users (u1, u2) annotating with different tags (t1, t2) the same movie (m). With dyadic relations, we could capture the user–movie, and the movie–tag relations, but we could not attribute tags to users (see Fig. 1)

Such and similar cases are best represented as a hypergraph. In a hyper-graph the edges, named hyper-edges comprise two, three or more nodes of potentially different type. An important issue is to discover the most important nodes in such a network; important nodes are of high *network value* and can be influential in the of information. For instance, in a folksonomy that is comprised of items, tags and users, a tag might be used often by many users to annotate different resources, and thus it might rise in prominence determining the influence of a new concept (consider for instance the use of the word google as a tag). Similar arguments can be drawn for influential users.

Figure 1: Triadic Relations and Hypergraphs



This work aims to study influence propagation in hypergraphs and in particular to characterise the structural properties of influential nodes. In infection propagation models we focus on the Susceptible—Infected—Susceptible (SIS), and on the Susceptible—Infected—Recovered (SIR) (Hethcote 2000). In the first model all nodes are initially in the susceptible state, then they can pass to an infected state, ending up in the susceptible state. In the SIR model, a node once infected can pass to the recovered state, where it can no longer be infected.

Intuitively, in single partite networks, the nodes at the core of the network are expected to be more influential than the ones at the periphery of the network. The k-shell decomposition maps the nodes of the network to cores of different values (or k values). The value of a core (or shell) denotes the level of cohesion among the nodes that belong to that core; thus nodes that belong to a k-core have at least k paths between them. The concept of cores was proposed in (Seidman 1983). The computational complexity of k-shell decomposition is linear to the size of a graph.

The current work employs the concept of k-shell decomposition to hyper-graphs, to assign k-shell values to nodes in order to discover whether the k-shell value is a more accurate predictor of the influence of a node as compared to degree centrality; we also consider the role of eigen vector centrality. Influence is measured as information propagation under the SIS and SIR models of infection.

Literature Review

The areas that are relevant to our research include the k-shell centrality, influence studies based on that measure, as well as centrality measures on hyper-graphs. Next, we report on those areas as related to this paper.

The definition of a k-shell (the term core is also in use) in simple graphs is as follows: A subgraph C of a graph $G = (V, E)$, where V and E represent the vertices and edges respectively, is a k-shell iff $\forall v \in V(C) : degree_G(v) \geq k$ and C is the maximum subraph with this property. cores were introduced in (Seidman 1983) as a measure of the cohesiveness of a network. An equivalent definition is that k-shell is a maximal subset of vertices such that each is reachable from each of the others by at least k vertex independent paths. Two paths are defined as vertex independent if they share none of the same vertices, with the exception of the start and the end vertex (Newman 2010, Sect.7.8.2). The computational complexity of the k-shell decomposition is $O(n + e)$, for a graph of n nodes, and e edges; this is an advantage of the k-shell centrality as compared to other centrality measures. The computation of the shell value for each vertex, i.e. the k-shell decomposition, starts with $k = 1$ and proceeds incrementally until all nodes have been assigned to a core.

Identifying the structural properties of influential nodes in a network has been studied in the case of undirected graphs under the SIS and SIR models of infection (Kitsak et al. 2010), with a constant probability of infection between neighbouring nodes. The following structural properties were studied: k-shell value, degree centrality, and betweeness centrality. It was discovered, on various real world networks, that infections originating from a single node of high k-shell value tend to spread the furthest compared to nodes of high centrality under the other measures.

The concept of k-shell has been extended to weighted graphs, by defining the weighted k' degree of a node as $k'_i = \sqrt{k_i \sum_j^{k_i} w_{ij}}$ and using it to derive *weighted cores* (Garas, Schwitzer, and Havlin 2012). Infection has been modeled according SIR, but the probability of a node to infect a neighbour is proportional to the weight of their connection. It was discovered, that an infection can be more widespread in weighted than in unweighted networks.

A more general approach aims to separate the core from the periphery of the network by allowing the discovery of cores of different sizes and shapes (Rombach et al. 2012), which extends earlier work (Borgatti and Everett 1999). In particular, for each node i the so named aggregate core value cs is computed as follows, $cs(i) = \sum_\gamma C_i(\gamma) \times R(\gamma)$, where $C_i(\gamma)$ expresses the local core quality of node i, $R(\gamma)$ expresses the global core quality, and γ is a parameter vector that defines the size and the shape of the core. The core quality is defined as $R(\gamma) = \sum_{i,j} A_i f(C_i, C_j, \gamma)$, where A is the adjacency matrix, and f is a function that can be defined accordingly. The point was to maximise the aggregate core value, which was performed with simulated annealing.

The concept of k-shell has also been introduced to graphs that evolve over time, with the aim to study the spread of an infection (Miorandi and Pellegrini 2010). In particular, two models have been proposed: the *flat* and the *rich* one. In the former, two nodes i and j are considered linked if they are connected for any time interval; in the latter model the duration of the connection plays an important role. Based on experimental evidence, it was discovered that if the contact duration of nodes follows a heavy-tail distribution then the k-shell value, as defined by the rich model, is an accurate predictor of the spread of infection compared to the flat model and compared to degree centrality. Relevant research has been carried out in the detection of k-shells in evolving graphs, without considering spread of infections; incremental algorithms have proposed by (Li and Yu 2012) with quadratic complexity, and in (Sariyuce et al. 2013) with linear complexity.

Hyper-graphs have been used to represent metabolic networks, food-webs, as well as in other domains. In hyper-graphs an edge, also known as hyper-edge, might comprise more than two nodes. The representation of hyper-graphs as simple networks is also possible, but some of the semantics are lost, which has as consequence that the degree centrality of nodes in a hyper-graph is potentially different from that of a simple graph, which in our case would affect the extraction of cores. In a study of human collaboration in paper authoring, it was discovered that the degree centrality of certain nodes depends on the representation form, thus a simple graph and a hyper-graph render different centralities to nodes (Estrada and Rodriguez-Velazquez 2005). A hyper-graph can also be represented as a bi-partite graph, where we introduce nodes to represent hyper-edges, this however increases the complexity of the graph by increasing the number of nodes.

A multipartite network is essentially a hyper-graph, that includes nodes of more than one type. The degree of a node is defined as the number of hyper-edges in which the node participates in. Such a graph can be represented with an incidence matrix, where rows represent hyper-edges, and the columns nodes. For instance Figure 2 represents a hyper-graph with edges appearing as curves, the corresponding incidence matrix appears in Table 1; in particular this is an example of a tri-partite network that might represent a social tagging system.

In hyper-graphs, the concept of node influence or centrality has been addressed by an extension of eigen vector centrality (Bonanich, Holdren, and Johnston 2004). Let $\mathbf{E}$, be the incidence matrix of a hyper-graph. Then the solution to the following eigenvector problem computes centrality scores x for the hyper-edges: $\mathbf{E}\mathbf{E}^\top x = \lambda^2 x$, here the centrality scores y for the nodes is computed as: $\mathbf{E}^\top \mathbf{E} y = \lambda^2 y$.

Folkrank (Hotho et al. 2006) is measure that ranks the nodes in a folksonomy that typically comprises users, tags and resources. Folkrank, is an extension of pagerank (Brin and Page 1998), which is based on concept of eigenvector centrality. The centrality of nodes denoted by $\vec{w}$ is defined as $\vec{w} = dA\vec{w} + (1 - d)\vec{p}$, where A is the adjacency matrix, $d \in (0, 1)$ is constant, and $\vec{p}$ is set to be a vector of all aces. Usually, the previous equation is repeated many times until the final value of $\vec{w}$ is obtained.

The previous two measures, i.e. eigenvector and folkrank centrality, model conservative diffusion process, which means that a quantity is preserved with the passage of time; this is closely related to the random walker model. On the other hand the initial quantity in a non-conservative process can increase or decrease. Current empirical evidence suggests that the diffusion of information is best described as a non-conservative process (Ghosh and Lerman 2010).

k-shell Decomposition for Hyper-graphs

Let V denote the set of vertices and E denote the set of hyper-edges, where each edge is a subset of V. Then $G(V, E)$ is a *hypegraph*. A vertex v is incident to an edge e, if $v \in e$. The degree d of vertex v can be defined as the number of edges that this vertex belongs to, same as in simple graphs. In particular, if the number of types of nodes is l, and each hyper-edge comprises l different nodes, the corresponding hyper-graph is named as a uniform l-partite graph or an $< l, l >$ graph for short. An edge e, being a set of nodes, is maximal if there is not another one that is a superset of it, that is $\not\exists e' : e \subset e'$. An edge sequence $P = (e_1, e_2, \ldots e_k)$ is named a path, provided that (Wang and Lee 1998):

$$\begin{aligned} \text{for } i \neq j, &\quad e_i \neq e_j, \\ \forall i, &\quad e_i \cap e_{i+1} \neq \emptyset, \\ \text{for } i \neq j, &\quad (e_i \cap e_{i+1}) \setminus (e_i \cap e_{i+1}) \neq 0, \end{aligned}$$

A k-shell contains all the nodes that have at least k vertex independent paths between them. In hyper-graphs, shells can be obtained much like as in simple graphs. That is first, all the nodes of degree one are removed, they are assigned to shell 1. Then edges with less than two nodes are removed, and the degree of the remaining nodes is modified accordingly. In the next step, nodes will be assigned to shell 2; and so on, until all nodes have been assigned to a shell (see also Algorithm 1). For instance, a k-shell decomposition, in two shells, of the graph in Figure 2 is depicted in Figure 3. The concept of discovering hyper-shells has also been addressed in computational biology in (Ramadan, Tarafdar, and Pothen 2004). In that work, in process of k-shell decomposition, edges which are not maximal are discarded. In our case, as we compute the shells, we do not discard non maximal edges, for they represent essentially extra connections that are used to spread information. Furthermore, in the process of k-shell decomposition we remove hyper-edges that contain only one node.

Let us now define the direct neighbours of a node u as all the nodes that belong to the same hyper-edge as u. This definition allows us to extend breadth first in hyper-graphs, which will in turn be used to implement infection models. In Figure 4 a breadth first traversal is depicted, of the hypergraph, that is represented in Figure 2 and Table 1. In that, we start from node u_1, the direct neighbours of u_1 are the t_1, r_1, r_2, t_4; the neighbours of the neighbours of u_1 are the u_2, and t_2, t_5, u_3.

Experiments

In the empirical evaluation we aimed to discover which of the degree, shell, or eigen centrality is the most accurate pre-

Table 1: Incidence matrix representation of a hyper-graph

edge	user	tag	resource
E_1	u_1	t_1	r_1
E_2	u_2	t_2	r_1
E_3	u_1	t_4	r_2
E_4	u_2	t_3	r_2
E_5	u_3	t_5	r_1

Figure 2: Hypergraph

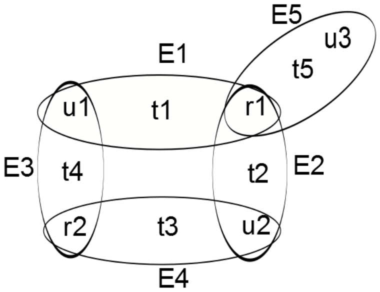


Figure 3: Shell decomposition of a hypergraph. All nodes are in shell-1, and the nodes in dark background are also in shell-2

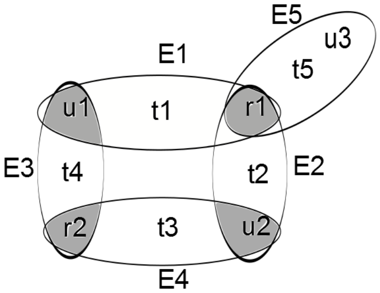
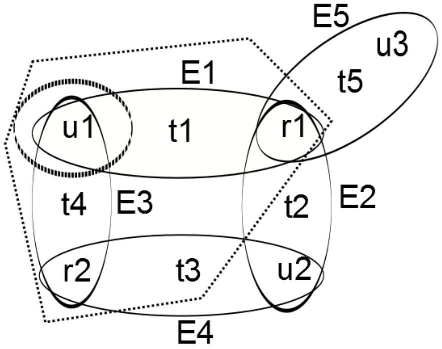


Figure 4: Breadth first traversal of a hypergraph



dictor of a node's capability to infect, under the SIS and SIR models.

In the experiments we used three data sets that were released as part of the HetRec workshop (Cantador, Brusilovsky, and Kuflik 2011); in particular the movie lens, delicious and lastFM sets. [1] Movie lens comprises: 2,113 users, 10,197 movies, 13,222 tags, 47,957 hyper-edges; LastFM comprises 1,892 users, 17,632 artists, 11,946 tags, 186,479 hyper-edges; and delicious comprises 1,867 users, 69,223 urls, 40,879 tags, 437,593 hyper-edges. All three data sets are uniform tri-partite networks, where the hyper-edges denote the assignment of a tag to a resource by a user.

First, we used the largest component in each of the tri-partite graphs. Second, we discovered the degree centrality, k-shell centrality and eigen centrality for each node in the tri-partite graphs. The results of k-shell decomposition appear in Table 2, where the maximum degrees and the number of shells are depicted. Moreover, it was discovered that low degree nodes almost coincide with nodes in outer k-shells as depicted in Figures 5,6 and 7. The y-axis denotes the percentage of nodes with identical k-shell, and degree.

Then we performed infections according to the SIS and SIR models having as starting point one node only. This was repeated for all nodes. The probability of a node infecting a neighbouring node, β was set to 5%. The SIS and SIR models appear in Algorithms 2, and 3 respectively. Infection emanates in discrete time steps in a synchronous way, for a maximum number of time steps *maxSteps*, which was set to 7. Second, each node once infected, remains infected for a certain time period $infec_t$, which was set to 2. After this, the node passes to the susceptible state in the SIS, or to the recovered state in the SIR models.

As a measure of comparison we used a loss function, that considers the extend to which high k-shell, degree or eigen centrality accounts for the most infectious cells. Next, p represents the percentage of the top p cells according to a centrality measure; e_{k_s}, e_d, e_{eig} are the loss functions for k-shell, degree and eigen centralities. In the following equation, the denominator denotes the top $p\%$ most infectious nodes irrespective of the centrality, whereas the nominator denotes the contagion capability of the top $p\%$ with respect to their centrality:

$$e_{k_s}(p) = 1 - \frac{M_{k_s}}{M_{\text{eff}}} \tag{1}$$

$$e_d(p) = 1 - \frac{M_{k_d}}{M_{\text{eff}}} \tag{2}$$

$$e_{eigen}(p) = 1 - \frac{M_{k_e}}{M_{\text{eff}}} \tag{3}$$

Each experiment was repeated 10 times, for $p \in [0.1 : 0.01 : 0.05]$ (in matlab notation). The results are depicted in Figures 8—13. The k-shell loss function is depicted as a solid line, degree loss function as a dash line, and the eigen loss function as a dot line. On the figures the error bars represent standard deviations across the 10 experiments. As it can be seen the k-shell loss function for different values of p, is a better predictor of the spreading capacity of a node; and it is correlated with the eigen loss function. This is an interesting result given that eigen vector centrality is more suited to conservative processes as discussed in the the literature review.

Table 2: Decomposition Analysis of Tri-partite graphs

Movie Lens		Delicious		Last FM	
degrees	cores	degrees	cores	degrees	cores
5430	640	5550	180	7503	421

Furthermore, in the case of movie lens, especially under the SIS model all three centrality measures performed poorly in detecting the majority of the most influential nodes. Without any further investigation we might speculate that another centrality measure, such the the in-betweeness centrality (Freeman 1977), might have a higher performance. Indeed, if the graph is made of multiple cores that are weakly connected, then a seed node inside a core might not take the infection very far.

In some experiments, not included in this report, we increased the value of $infect_t$, gradually up to $maxSteps$. The result was that the loss functions remained well separated for both the degree and the k-shell centrality in the SIS, and SIR experiments; but the loss functions have higher values. This is caused by older nodes, which retain the capability to infect. Moreover, in another series of experiments we increase the value of $maxSteps$ up to 3 times the value we set in the current experiments. The result is that the loss function for degree centrality tends to become indistinguishable from the k-shell function. Finally, beyond a certain threshold of the infection probability (i.e. around 20%), the three infection methods under both the SIS, and SIR produce almost identical results.

Algorithm 1: Hypergraph k-shell decomposition

```
$V$: hypergraph nodes;
$E$: hypergraph edges;
$i \to 1$;
while $V \neq \{\}$ do
    while $\exists u \in V$ with $degree(u) = i$ do
        remove $u$ ;
        if $\exists e \in E$, with one node only, then
            remove $e$;
    $i \to 1$
```

Conclusions and Future work

The experimental study on three hyper-graphs that represent social networks confirms that the k-shell centrality in is a more accurate predictor of influence than degree centrality under the SIS, and SIR model. Thus we extend similar studies on simple graphs. The eigen vector centrality seems to be similar, in predicting influential nodes, to k-shell centrality.

[1] All data sets are hosted in the group lens site `http://www.grouplens.org/node/462`

Algorithm 2: SIS Infection in hyper-graphs

```
$I \leftarrow \{\}$ :infected nodes ;
$V$ :graph nodes;
$s \in V$: Starting node of infection;
$I \leftarrow I \cup \{s\}$ ;
for i=1 to maxSteps do
    $\forall v \in I$ infect their direct neighbours with probability $\beta$;
    let $T$ be the infected neighbours, $\forall u \in T$, set $u.life = infect_t$;
    $I \leftarrow I \cup T$;
    $\forall v \in I$, $v$.life $\leftarrow$ $v$.life$-1$;
    $\forall v \in I$, where $v$.life$= 0$, $I \leftarrow I - \{v\}$;
```

Algorithm 3: SIR Infection in hyper-graphs

```
$I \leftarrow \{\}$ :infected nodes ;
$R \leftarrow \{\}$ :recovered nodes;
$V$ :graph nodes;
$s \in V$: Starting node of infection;
$I \leftarrow I \cup \{s\}$ ;
for i=1 to maxSteps do
    $\forall v \in I$ infect their direct neighbours who are not in $R$ with probability $\beta$;
    let $T$ be the infected neighbours, $\forall u \in T$, set $u.life = infect_t$;
    $I = I \cup T$;
    $\forall v \in I$, $v$.life$\leftarrow$ $v$.life$-1$;
    $\forall v \in I$, where $v$.life$= 0$, $I \leftarrow I - \{v\}$ and $R \leftarrow R \cup \{v\}$
```

We should also note, that k-shell as a centrality measure can be computed in time linear to size of hyper-graph, whereas other measures are computationally more demanding. For instance the folkrank measure, requires time linear to the size of the graph for each iteration.

Moreover, in the future we intend to study the role of cores in influence propagation in evolving hyper-graphs. On line social networks evolve by the addition or deletion of nodes and edges. There has been some work on assigning core values to each node in evolving graphs (see (Miorandi and Pellegrini 2010), (Li and Yu 2012), (Sariyuce et al. 2013)); this will be extended to hyper-graphs and then influence propagation can be studied.

Finally, in the current study, we computed centralities and studied influence irrespective of the types of nodes. We can advance beyond that, by redefining centrality measures to account for multiple node types (see for instance the work in (Becker 2013)), and based on that to reconsider influence modeling.

Acknowledgements

The author is grateful to the American College of Greece, “Deree” and to the Institute of Informatics and Telecommunications of the NCSR “Demokritos” for their support. This work is related to the research project REVEAL (REVEALing hidden concepts in Social Media), which is funded by the European Commission, under the FP7 programme. Moreover, the author would like to thank the anonymous reviewers for their constructive comments.

References

Becker, N. 2013. Ranking on multipartite graphs. Diploma thesis, Institute of Computer Science, LMU, Munich.

Bonanich, P.; Holdren, A. C.; and Johnston, M. 2004. Hyper-edges and multidimensional centrality. *Social Networks* 26:189–203.

Borgatti, S. P., and Everett, M. G. 1999. Models of core / periphery structures. *Social Networks* 21:375–395.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference (WWW 1998)*, volume 30, 107–117.

Cantador, I.; Brusilovsky, P.; and Kuflik, T. 2011. 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011). In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys 2011. New York, NY, USA: ACM.

Estrada, E., and Rodriguez-Velazquez, J. 2005. Complex Networks as Hypergraphs. *arXiv:physics/0505137*.

Freeman, L. 1977. A Set of Measures of Centrality Based on Betweenness. *Sociometry* 40(1):35–41.

Garas, A.; Schwitzer, F.; and Havlin, S. 2012. A k-shell decomposition method for weighted networks. *arXiv:1205:3720v1*.

Ghosh, R., and Lerman, K. 2010. Predicting influential users in online social networks. *CoRR* abs/1005.4882.

Hethcote, H. W. 2000. The mathematics of infectious diseases. *SIAM Rev.* 42(4):599–653.

Hotho, A.; Jäschke, R.; Schmitz, C.; and Stumme, G. 2006. Folkrank: A ranking algorithm for folksonomies. In *University of Hildesheim, Institute of Computer Science*, 111–114.

Kitsak, M.; Gallos, L. K.; Havlin, S.; Liljeros, F.; L., M.; Stanley, H. E.; and Makse, H. A. 2010. Identifying influential spreaders in complex networks. *Nature Physics* 6:888–903.

Li, R., and Yu, J. X. 2012. Efficient core maintenance in large dynamic graphs. *CoRR* abs/1207.4567.

Miorandi, D., and Pellegrini, F. D. 2010. K-Shell Decomposition for Dynamic Complex Networks. In *Proceedings of the International Workshop on Dynamic Networks (WDN)*.

Newman, M. E. J. 2010. *Networks: An Introduction*. Oxford University Press.

Ramadan, E.; Tarafdar, A.; and Pothen, A. 2004. A hypergraph model for the yeast protein complex network. In *Proceedings of the Sixth IEEE Workshop on High Performance Computational Biology*.

Rombach, M. P.; Porter, M. A.; Fowler, J. H.; and Mucha, P. J. 2012. Core-Periphery Structure in Networks. *arXiv:1202.2684v2*.

Sariyuce, A. E.; Gedik, B.; Jacques-Silva, G.; Wu, K. L.; and Catalyurek, U. V. 2013. Streaming algorithms for k-core decomposition. In *International Conference on Very Large Data Bases (VLDB)*.

Seidman, S. 1983. Network structure and minimum degree. *Social Networks* 5:269–287.

Wang, J., and Lee, T. T. 1998. Paths and cycles of hypergraphs. *Science in China* 42(1).

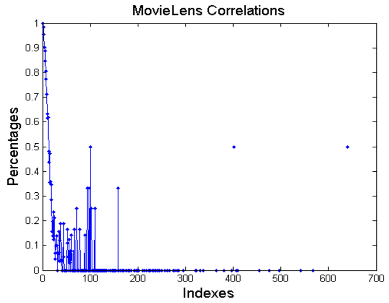


Figure 5: Movie Lens: Degree and Shell Correlation

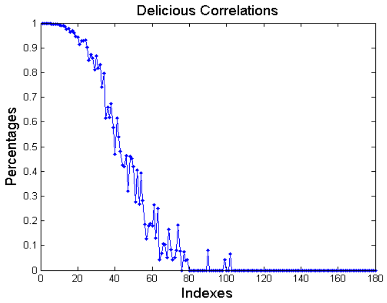


Figure 6: Delicious: Degree and Shell Correlation

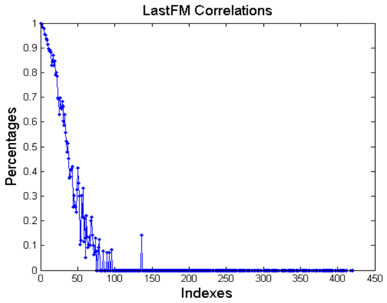


Figure 7: Last FM: Degree and Shell Correlation

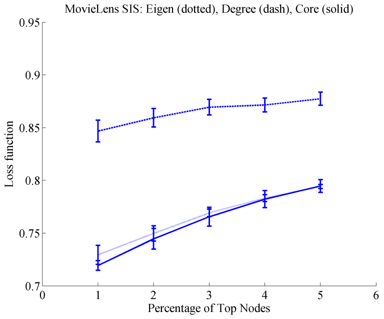


Figure 8: MovieLens SIS, Infection prob.=5%

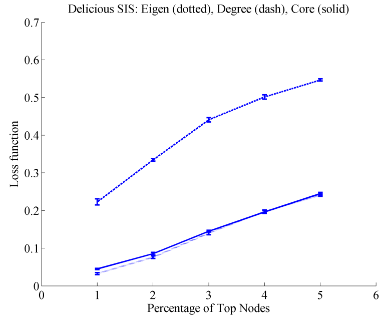


Figure 9: Delicious SIS, Infection prob.=5%

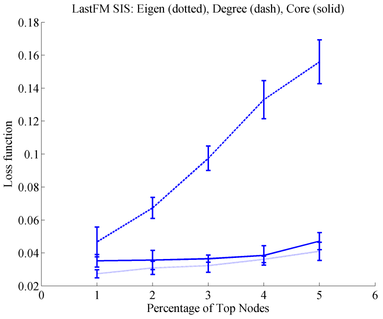


Figure 10: LastFM SIS, Infection prob.=5%

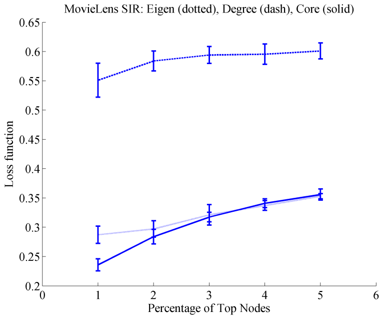


Figure 11: MovieLens SIR, Infection prob.=5%

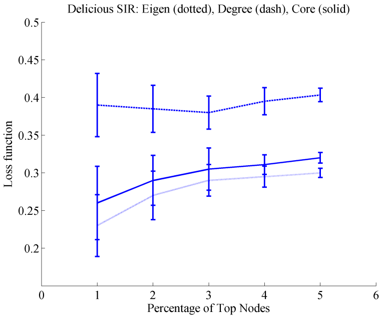


Figure 12: Delicious SIR, Infection prob.=5%

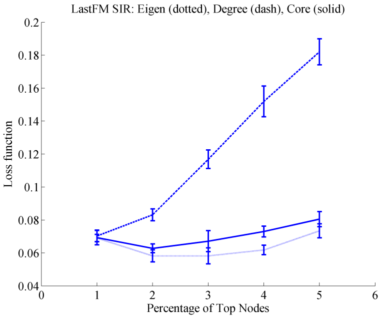


Figure 13: LastFM SIR, Infection prob.=5%