Working for influence: effect of network density and modularity on diffusion in networks

Habiba
Department of Computer Science
University of Illinois at Chicago
Chicago, USA
hhabib3@uic.edu

Tanya Berger-Wolf
Department of Computer Science
University of Illinois at Chicago
Chicago, USA
tanyabw@uic.edu

***Abstract*—The problem of finding the most influential individuals, or the largest spreaders, in networks has been shown to be NP-complete even for simple spreading models, though approximable by a simple greedy algorithm. Yet, even the greedy algorithm relies on stochastic simulations that can be quite time consuming and intractable for large networks. Recently developed heuristics are fast and work well in practice but are limited to certain network models, spreading goals, or sampled networks. In this work, instead of devising a new spread optimization method, we re-examine the problem by analyzing the global structural properties of the underlying network as indicators of spread trends. Specifically, our investigations use density of a network as an indicator of: (a) when it is necessary to employ a sophisticated yet computationally expensive method? or (b) when even a random set of spread initiators perform as well as the best in expectation for maximizing the spread in the network? and (c) why certain heuristics like high degree as indicator of high spread work for certain networks and not for others. We show that for network densities above and below a certain threshold, the difference between the best and expected spread is negligible. In between the two extremes, the networks exhibit marked differences between the best and expected spread. This region, rich with non-trivial and complicated structures, requires further work to devise efficient techniques for finding best spreaders.**

***Keywords*-diffusion, network structure, density, block-mixture model**

I. Introduction

Who are the most influential individuals in a (social) network and can we efficiently find them? This has been the subject of extensive research in applications of social sciences, economics, viral marketing, and epidemiology, among others [5, 7, 9–11, 14, 17, 20, 21, 24, 26, 27, 30, 31, 33, 35, 37, 42]. Unfortunately, computationally this and similar questions have been shown to be in the class of NP-hard problems. Moreover, even the best approximation algorithms are infeasible for massive network data available on todays ubiquitous large platforms, such as social networking websites, blogosphere, and communication networks. However, is it really necessary to expend this intensive computational effort in order to identify those individuals? While theoretically the answer seems yes, in practice many efficient heuristics have been demonstrated to be effective. In this paper we show that the hardness of computationally finding the most influential individuals in a network, depends on the structural properties of that network, such as its density (or, rather, the effective density, which is the combination of density and the probability of infection, that is, the density within the spread network) and modular structure.

The problem of identifying influential individuals in a social network has been most rigorously formulated and analyzed by Kempe et al. [26]. The problem has been shown to be NP-complete and a greedy approximation algorithm guarantees a solution no worse than $(1 - 1/e)$ factor of the optimal for many general models of the spread of influence. To find the set of k most influential individuals, the greedy algorithm computes the most influential individual and removes it along with those it influences from the network and repeats the process. However, to find this one most influential individual at every iteration the algorithm relies on multiple repetitions (typically thousands) of stochastic simulations of the process of spreading the influence through the network. The larger the network, the many more iterations are necessary for each stochastic simulation. Moreover, a separate set of simulations is necessary for each set of parameters of the influence spreading model. This process becomes infeasible even for moderately sized networks of several thousands nodes. To date, no algorithmic or analytical way has been found to replace the stochastic simulations. Instead, many simple yet effective heuristics for finding influential individuals have been proposed that exploit structural network properties.

It has been shown that removing nodes (or individuals) of the highest eigenvalues reduces the spread (or influence) the most [36, 38, 40]. Expansion factor of a graph has been shown to be a good indicator for detecting the spread of a virus over the Internet [1]. The core versus periphery structure of the network has been shown to facilitate the inference of network of influence [18]. Immunization of hubs in a network [23] is a good strategy to minimize spread but clustering co-efficient is not an effective measure [22, 23] for such goals, or indirectly, immunizing acquaintances of randomly selected nodes can contain the low threshold spreads quickly [8]. Degree centrality heuristic has been studied extensively for such goals. For example, it has been shown to work well for containing spread of misinformation [6], shown to strongly correlate with rates of infection [25], and many other similar problems. In social sciences, as early as 1960's, Granovettors insight about the

strength of weak ties [19], and a large body of research on the diffusion of innovations [37, 39], has shown that an individual's ideas and behaviors are direct functions of the ideas and behaviors of the people the individual is connected to. However, no systematic, empirical or theoretical, study has been done to explain why, seemingly in contradiction to the theoretical results, and when do these heuristics work well. And, more importantly, what is it in the structure of the real social networks that allows them to perform so well.

The goal of this work is to find clues in the global network structure that make a particular heuristic for identifying influential individuals work better than others and to figure out when heuristics work at all and when a serious computational effort is unavoidable. Thus, we take a step back from devising yet another method that works for a certain set of networks, to answer a more general question of what makes a certain heuristic effective, and not another, for a given network? Specifically, we show that it is possible to use effective density and modular structure of a network as indicators of when it is necessary to employ a sophisticated yet computationally expensive method versus when even a random set of spread initiators will perform as well as the best, in expectation, for maximizing the spread of influence in the network. We find that the effective density, which is the product of the density and the probability of spread, is a better indicator of the extent of spread than the density on its own. A spreading process with very low spreading probability in denser networks behaves similar to the spreading process in sparse networks. We show that networks with (effective) densities above and below a certain threshold are amenable to simple heuristics. In dense networks, in fact, there is no differentiation between influence of individuals and any random set of individuals is good, given a high enough rate of spread. In effectively sparse networks simple heuristics like highest degree nodes perform well. In between the two extremes, the difference between the best and a random set of individuals may be significant. That difference, in fact, depends on how modular (non uniform) the network is. The more modular, non-uniform, the network is, the bigger that difference is. Thus, it is for those intermediate density networks with rich complicated structure that we need to use sophisticated and computationally intensive approaches to find influential individuals. This result supports the findings that simple heuristics perform well in practice since most real world networks are very sparse, with few exceptions that are very dense (that represent small single communities).

To demonstrate our results, we systematically empirically evaluate the difference between the expected and the approximate optimal extent of influence spread on a variety of synthetic and real world networks, over a range of generative models, densities, and degrees of modularity. Our results consistently tell the same story: density and modularity matter and can tell us when to use a simple heuristic and when to put the effort and use the approximation algorithms based on stochastic simulations. Indeed, once stated this way, it is not surprising and quite intuitive: in dense networks everybody is connected to almost everybody and any set of individuals will do well; and in sparse networks nothing spreads well beyond the immediate neighbors so high degree nodes do best. However, our work for the first time test this assertion in a systematic way, quantifies the critical network properties, identifying the transition thresholds and laying grounds for a more rigorous theoretical analysis.

II. Methodology

We simulate spread of influence in networks of varying densities and modular structure, using optimal, random, and heuristic criteria for selecting spread initiators. We record the expected extent of the resulting spread and compare the optimal to all other choices.

Our experimental analysis is based on a large set of real world networks as well as synthetic networks generated using well studied generative network models. For real world networks, the biggest challenge was to find a variety of networks that cover the entire range of densities. Most real world networks, such as, social networks, co-authorship networks, or the web, are very sparse. On the other hand, networks observed for studies like behavior or disease spread are generally *proximity* networks that are generally very dense. In this paper, we report results of two real world networks which we could sample for a range from low to high density by aggregating it over increasing time window size. For synthetic networks, we used the preferential attachment model to generate networks that resemble many real world networks. For finer insights into how network structure, other than its density, contributes to the disparity in optimal and expected or heuristic spreads, we use the block mixture model [34] to generate another set of synthetic networks. For each network, real or synthetic, we simulated a spreading process, over a range of parameters of the spreading model and compared the best, expected, and heuristic extents of the resulting spread in the network.

A. Real world networks

We have examined numerous real world networks including blogosphere, on-line social networks, email exchange networks, router networks, co-authorship, human proximity networks, and animal proximity networks[1]. The list of networks and their basic statistics is given in Table I. The datasets statistics in Table I show that most real world networks are very sparse and the results of our experiments are consistent with all other results on sparse networks shown in this paper. Two of these networks are dense: Reality Mining and Plains zebra. These networks are also dynamic and by varying the length of the window over which the network was aggregated we could sample the network at different densities.

[1]Complete results and references of these datasets are available at: http://compbio.cs.uic.edu/∼habiba/diffusion–networkStructure.html

Table I: Real networks statistics

Dataset	Nodes	Edges	Density
AS	16299	34157	.0002570
DBLP	964	1891	.0004074
Live Journal	15001	66286	.0005890
P2P	8114	26013	.0007903
Epinions	9997	216213	.0043270
Political Blogs	1224	16715	.0223320
Pol Books	105	441	.0809520
Karate	34	75	.1336900
Enron	147	3467	.3229890
Plains	282	29050	0.730606
Reality Mining	96	3625	0.794956
Onagers	29	402	.9901000
Grevys	28	779	1

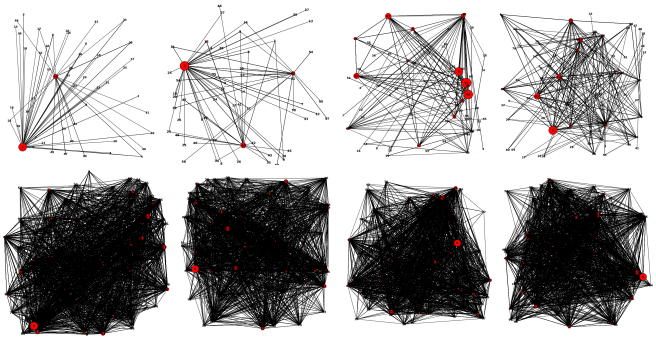
1) MIT Reality mining: The Reality Mining project represents the proximity network of Bluetooth devices of about 100 students and faculty members at MIT over 9 months. This is one of the first such systematic collection of data about human interaction and behavior [12]. We sample this network by aggregating the proximity data in one month size windows [2]. As is expected the density of interactions (proximity) increases with the each additional month. As a result, we get a series of sparse to dense networks for the Reality mining network of 9 months. Figure 1(a) represents a few samples of this network from low to high densities.

2) Plains zebra network: Social interactions between Plains zebra (Equus burchelli) in Kenya were recorded by direct observations made by behavioral ecologists from Princeton University. The data was made from visual scans of the population, typically once a day, over a period of several months. Each entity is a zebra, uniquely identified by the pattern of its stripes. Each spatially proximate group of animals, as determined by GPS coordinates, represents a complete set of interactions amongst those individuals [15]. Similar to our experiments on the Reality Mining dataset, we aggregated fixed size samples of the network overtime. We generated a time series of sparse to dense networks of Plains zebra. A sample of the series of increasing density Plains networks is given in Figure 1(b).

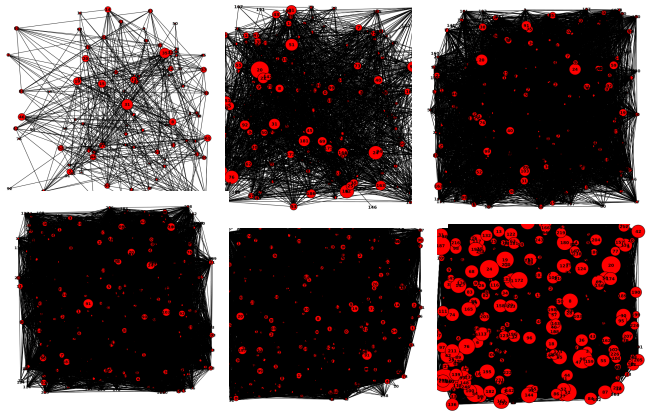
B. Synthetic networks

To rigorously test our conjecture in controlled parametric settings we used generative network models that have been shown to represent real world networks closely. We generated synthetic networks using the extensively studied Preferential Attachment(PA) model for scale-free(SF) networks [28]. This model, among other properties, represents the skewness of the degree distribution of many real world networks. We also used the stochastic block mixture model [34]. This model highlights the modularity of many real world networks (especially biological networks).

[2]The size of the appropriate aggregate window for network sampling is in general a nontrivial problem and is beyond the scope of this paper.



(a) Reality Mining



(b) Plains zebras

Figure 1: Series of progressively (successively) accumulated networks. The size of the node represents the relative degree of the node.

1) Preferential attachment model for scale-free networks: The web, citation networks, and the network of film actors are among a few examples of social networks that have been shown to exhibit a skewed degree distribution [2–4, 41] . Preferential attachment model [4] generates networks with skewed degree (specifically, power law) distributions. This is one of the first network generative models. In this model the network evolves over time. Nodes are added to the network sequentially, each new node creates links to already existing nodes proportional to how well connected the other nodes are. Hence, a new node is much more likely to get connected to a high degree node than to a low degree node. The idea is based on the premise of "rich getting richer". Many real world networks have been shown to exhibit this process of growth. The skewness of the degree distribution is determined by a parameter γ which is usually set between 2 and 3 based on the type of network being studied. We generated networks of more than 30 different densities ranging from very sparse to very dense networks. Each density is sampled 10 times.

2) Block mixture model: The Erdös-Rènyi (ER) random graph [13] model is simple and is completely defined by one parameter, namely the probability of an edge. In this model, given a fixed set of nodes and a probability p,

each edge is generated independently uniformly at random with that probability p. The resulting graph, of course, has density $p\binom{N}{2}$. The model's many asymptotic properties have been well studied which makes it ideal for analyzing the relation between density and the extent of spread in a network. However, the model does not fit well to real world networks. For example, nodes of real world networks are often structured in tight relatively well-connected clusters (communities) not captured by ER model. The stochastic block mixture model was proposed for this purpose in the context of social sciences, using a Bayesian approach [34]. Further refinements, such as the assortative mixing [29] has made this model a natural choice to analyze real world networks in controlled parameter settings. The block mixture model designates the nodes into C blocks. Given two parameters, the inter– and intra–block edge probabilities, the edges are generated uniformly at random, with the appropriate probabilities for each pair of nodes. We use this model to generate networks of very low to high densities while varying the inter– and intra–block probabilities. The resulting set of synthetic networks not only provides us with data to compare density with the extent of spreads but also gives us better insight into the modular structure of networks to further refine our analysis.

C. Spreading process

We simulate spread in the real world and synthetic networks using the Independent cascade spreading model. This model was first introduced in the context of word-of-mouth marketing [10, 16]. This is also the most commonly used simple model to study disease transmission in networks [9, 30, 32, 33, 35]. In this model, transmission from one individual to another happens independent of interactions with all the other individuals. This model describes a spreading process in terms of two types of individuals, active and inactive. The spreading process unfolds in discrete timesteps. In each timestep, each active individual attempts to activate each of its inactive neighbors. The activation of each inactive neighbor is determined by a probability of success. If an active individual succeeds in affecting any of its neighbors, those neighbors become active in the next time step. Each attempt of activation is independent of all previous attempts as well as the attempts of any other active individual to activate a common neighbor.

D. Extent of spread simulations

The extent of spread in a network is the number of individuals affected at the end of a spread process (simulation), initiated by a set of individuals. Here we measure this extent based on three types of spread initiators.

1) Optimal spreaders: The k spreaders that maximize the extent of spread in the network. Since for large networks it is hard to find the exact k best spreaders, we use the greedy approximation of Kempe et al [26] for $k > 1$.
2) Random spreaders: Any k spread initiators picked uniformly at random from the network.
3) Ad–hoc spreaders: k top ranked individuals based on some network property that are used for initiating spread. In this work we used degree, eigenvalue, and boundary nodes to rank nodes in the network.

For each network we find the extent of spread using the three types of spreaders. We compare the maximum spread in the network - based on the optimal spreaders to the expected spread in the network - based on randomly selected sample of individuals. We also compare the optimal spreaders to the ad hoc spreaders. The difference in the extent of spread, for each pair of methods, highlights the disparity in the network structure and how it affects the resulting extent of spread. We focus on how this difference varies with network density. All spreads simulations are based on the Independent cascade model described in the previous section.

III. Results and analysis

Partial aggregations of a dynamic network over a set of individuals give a series of networks of increasing density over those individuals. This process of adding (but not removing) interactions to create increasingly denser networks is similar to the way most network generative models are defined. Moreover, assuming the underlying generative process which is evidenced by the network is stable, those series represent networks with similar dynamics. As a result, we have a set of networks from the same domain and interaction dynamics but varying (increasing) densities. Figure 1 shows the series of aggregated Reality Mining and Plains zebra networks. Clearly, the two networks become denser as more data are sampled and added to the network. Similarly, we generate a series of sparse to dense Scale–free networks using the Preferential Attachment model [4].

Recall, that for each network in the series we have simulated a spreading process (with many repetitions), using each node, in turn, as the spread initiator. We have computed the resulting extent of spread and can now compare the optimal, the expected (mean) and the heuristic–based resulting spreads. We observe the following three trends in our experimental analysis. (a) The optimal extent of spread is well approximated by the expected extent of spread in networks with very low or high densities but not in the mid-range density; (b) degree–based heuristics result in remarkably near–optimal extent of spread in lower density networks, but these heuristics may not perform consistently well for denser networks; And (c) modularity of a network can be exploited for designing better heuristics that approximate the optimal spread well especially for networks of mid-range density.

A. Optimal versus expected spread

Figures 2(a), 2(b), and 2(c) show the difference between the optimal spread and the expected spread as a function of the effective density in Reality Mining, Plains zebra networks, synthetic networks respectively. The plots show the following three trends in the extent of spreads as the

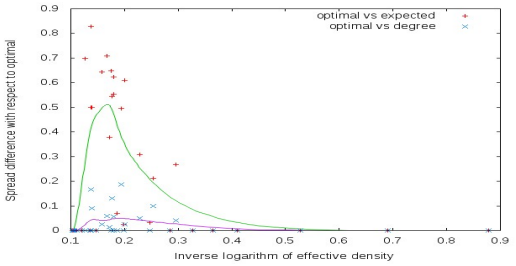
effective densities (density $\times$ probability q of influence) of the networks progressively increase.

First, at very low effective densities ($\leq$ .004 for real networks and $\leq$ .001 for synthetic networks), the extent of spread is very low, irrespective of how sophisticated the approach for selecting spread initiators is. Low density networks lack a "well defined" structure, by definition, most nodes in low density networks are sparsely connected. Such low density networks have extremely skewed degree distribution: a very small number of disproportionately high degree nodes and the remainder of sparsely connected nodes. Hence, in these networks, the extent of spread is very low relative to the size of the network even with high probabilities q of influence (which is essentially a deterministic spread), since most nodes have few or no neighbors to whom they can propagate the spread. Only high–weighted degree nodes can influence many others but there are so few of them, and their neighbors have such low degrees, that they hardly make a dent in the overall extent of spread.

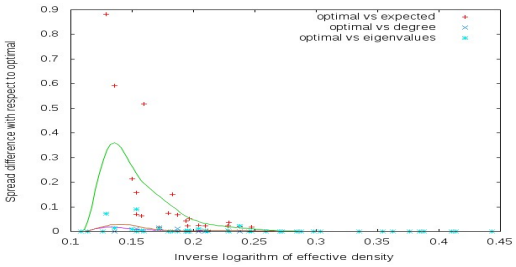
The second trend we observe is that at high effective densities (≥ 0.25 for real networks and $\geq .0035$ for synthetic networks), most nodes are uniformly well connected. In expectation, the extent of spread initiated by any random node is high and is comparable to the optimal spread in these dense networks due to the similarity in the connectivity of nodes. In networks of high densities, we find that spreading processes behave almost deterministically, that is, spreading process affect the entire connected component irrespective of who initiates it. Hence, similar to low density trends, we find that in such networks, the optimal approach does not outperform the expected by much.

Finally, and most interestingly, in networks of intermediate effective densities we find a clear phase transition in the difference between the extent of spread resulting from the optimal and random initiators. This difference in the optimal and the expected spread increases until it peaks and starts to decline gradually, as network densities progressively increases. In this intermediate region, the extent of spread is very sensitive to the identity of the initiator. Clearly, in this region it is worth while looking for an optimal or near–optimal spread initiator. Moreover, the structure of the network plays a role in how influential different nodes are. We further focus on this intermediate region in Section III-C to investigate how the edge topology affects the difference in the extent of spread from various nodes.

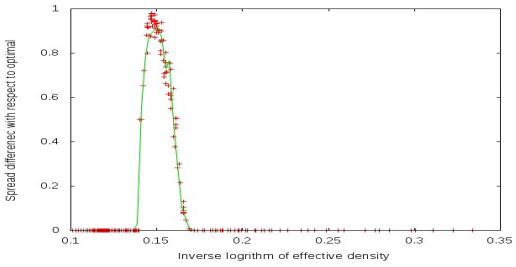
Note one difference between the Scale–free and the real world networks: compared to the real world networks, the high optimal versus expected difference region is confined to a smaller range of densities. Clearly, network characteristics other than just the density and the skewness of the degrees implicitly contribute to the extent of spreads that are not captured by the Scale–free networks. Before we delve further into this issue we analyze the behavior of some of the most well studied heuristics for influence maximization in the same effective density settings.



(a) Reality Mining



(b) Plains zebras



(c) Scale-free networks

Figure 2: The difference between the optimal and the expected. Figures (a-b) also show the difference between the optimal and the degree heuristic (cross signs) and figure (b) shows the difference between the optimal and the eigenvalue heuristic (asterisks). Spline function smoothing applied to draw approximation curves. X-axis are the inverse log of the effective density. Y-axis are the the relative difference w.r.t optimal spread.

B. Optimal versus heuristic spread

We compared both the optimal and the expected spread to the performance of some of the most well studied heuristics for influence maximization problem. Degree centrality and highest eigenvalues have frequently been shown to correlate with the influence maximization objective [6, 25, 36, 38, 40]. We simulated spread in Reality Mining and Plains zebra networks by using k highest degree and k largest eigenvalues nodes as spread initiators. We compared these results with the optimal and the expected extent of spread. The difference in the optimal and the heuristic extent of spread for $k = 1$ are shown in Figure 2(a) and 2(b) for Reality Mining and Plains zebra networks, respectively. For other values of k the trends are the same and we omit them here due to space constraints. We find that both heuristics give close to optimal solutions at extreme densities. In fact, at low densities the optimal spread initiators are indeed the nodes with the highest degree and highest eigenvalues. Thus, at low densities our results support what has been found in many other studies.

At high density, even though results from heuristics are comparable to the optimal, the identities of the optimal spread initiators and the heuristically chosen nodes are not the same. Recall, that at those densities any randomly chosen initiator performs as well as the optimal. Thus, the good performance of the heuristics here is not due to the identity of the chosen initiator but due to the fact that any node serves well. This eliminates the significance of a heuristic or even the optimal approach over a random one at high density. In the intermediate region, although the heuristics work remarkably better than random, the differences in the extent of spreads between the optimal and heuristic methods are inconsistent. Thus, a heuristic may work for some networks but not others, and there are no theoretical or empirical guarantees for its performance in this region. This trend has not been taken into consideration in formulating better methods for spread optimization. This study gives us a better understanding of the effects of network structure on spread and it shows where we need to put our efforts to exploit the structure of the intermediate densities for formulating better methods for the spread optimization problem.

For better insight into the effect of network structure on the extent of spread we analyze extent of spreads in synthetic networks generated using stochastic block mixture models.

C. Network modularity and extent of spread

Recall from Section II-B2, that a network can be represented as a set of components, or blocks, with *inter*– and *intra*–block probabilities for connectivity. For a well–defined clustering of nodes to exist it is assumed that the inter–block probabilities of forming links are much lower than the intra–block probabilities. Given this explicit modular structure of a network, intuitively a good strategy for influence maximization in such a network is to distribute the k spread initiating nodes among the blocks of the network, proportionally to the size of each block. This is exactly the expected distribution of the initiators if we select them uniformly at random from among all the nodes in the network. We compare the expected extent of spread resulting from such random initiator selection to the optimal (greedy) initiator selection, as we did previously. We observe essentially the same three trends in the extent of expected spread as we generate sparse to dense network using the stochastic block model. The control of the modular structure of the network, in addition to its density, supplements the observations made in earlier real world as well as synthetic networks analysis. The following are the trends we observe in the stochastic block model networks. First, in a sparse network, the probability of edges within a block is not much higher than across the blocks (otherwise the networks rapidly become too dense). The resulting network is therefore not much different from a sparse uniform random graph. Thus, as before, the optimal influence maximization strategy is not significantly more effective compared to selecting the top k highest degree nodes as spread initiators. Second, in very dense networks, similarly, the probability of connections within a block is high but so is the probability of connections across blocks, to make up for the very high density overall of the network. Again, this structure results in a network similar to uniform random graph and the extent of spread is the same (and large), regardless of the initiators. Thus, the best approaches are comparable to the random choices of spread initiators in dense stochastic block networks, as before. Third, in between the two cases is the intermediate density region where, in terms of stochastic block mixture model, blocks are clearly defined by the large differences in the inter– and intra–block probabilities of connections. Here our analysis is further refined into two sub-cases: In this intermediate density region the extent of spread using the optimal strategy differs significantly from the expected spread, where the spread initiators are chosen uniformly at random. This result supports what we have shown in real and scale-free networks in Section III-A. The choice of the stochastic block mixture model provides us with a better understanding of the effect of network structure on the extent of spread. Block mixture model is completely defined by the inter– and intra–block probabilities. Regardless of our knowledge of the modular structure of the network, the optimal influence maximization strategy would still pick the best of the $\binom{n}{k}$ possible choices as spread initiators. The brute force enumeration, as is the case with all large real world networks, is an infeasible strategy. But given that the structure of the underlying network is generated by the stochastic block mixture model, one intuitive influence maximization heuristic is to distribute the k spread initiators among various blocks proportionally to the blocks sizes. In this case, the expected spread in the network is approximately the sum of the extent of spread in each block and the sum of the extent of spread across the blocks. Let each block B_i have inter–block connection probability of p_i (for simplicity, we assume that to be uniform within a block). And let p_{ij} be the intra–block connection probability between any two blocks B_i and B_j. Let k be the number of spread initiators and q the probability of influence. Then, the expected spread $E(G, q, k)$ in the network $G(N, E)$, when the initiators are chosen uniformly at random from among all the nodes, is:

$$E(G = \{\cup_i B_i\}, q, k) \approx \sum_i E(B_i, p_i) + \sum_{ij} E(B_i, B_j, p_{ij}).$$

An improvement on the extent of the expected spread would be to choose the spread initiator by taking into consideration the modular structure of the network. One possible heuristic is to select nodes on the boundary of blocks, that is, nodes that connect one block to another. Such nodes are structurally well placed to be effective spreaders within and across blocks. It is also worth noting that such peripheral nodes have above average degrees. Since the expected degree of a node within the block B_i is p_i, the boundary node by definition has an expected degree of $p_i + \sum_{j \in \{\cup B_j\}} p_{ij}$. Thus, with relatively high degrees such nodes are better candidates for maximizing spread. In expectation, the spread

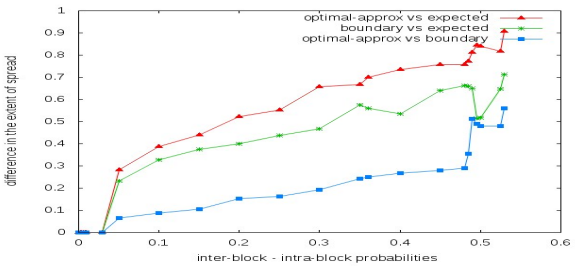


Figure 3: The difference between the Greedy approximation and the expected spread, the boundary heuristic and the expected spread, the approximation and the boundary heuristic spread.

initiated by the boundary nodes is greater than the expected uniformly initiated spread: $B(G,q,k) \geq E(G,q,k)$ and the optimal spread is, of course, at least that of the heuristic spread: $O(G,q,k) \geq B(G,q,k)$. Given that finding the optimal solution is not feasible (since the only known way is to use brute force enumeration), we can either approximate it again using the greedy approximation algorithm or use the difference between the expected and boundary spread as the lower bound on the difference between the expected and the optimal. Moreover, the boundary choice heuristic shows how modular structure of the network can be used to find a better influence maximization solution. Overall, we analyzed three types of spread differences in networks that were generated using the stochastic block model. We compared differences in the extent of the optimal and the boundary-node based heuristic spread, the boundary heuristic and the uniform expected spread, and the optimal and the expected spread for networks of various densities and inter– and intra–block probabilities of connectivity. Figure 3 shows that with the increase in network modularity, all the differences between the optimal and the expected, the optimal and the boundary heuristic, as well as the boundary and the expected spread increase as well. We observe that, in the intermediate density region, the difference between the inter– and intra–block connectivity is the highest. Within this high difference region, we find that simple degree–based heuristics, that take into account the modular structure of the network, perform much better than the expected and much closer to the optimal spread. Yet, this heuristic is much easier to compute than the optimal solution. Thus, the additional knowledge of the modular structure was helpful in devising a better heuristic for this specific model. We conclude that the effective density and modularity of a network are the two key properties that affect the extent of spread in the network. Consequently, these network properties should be incorporated in the design of efficient methods for influence maximization problem.

IV. Conclusions

This work is a systematic exploration of the connection between network structure, specifically, density (and the probability of influence), modularity, and extent of spreads in those networks. Our results show strong density effect on the extent of spread and the ease of finding influential individuals. Extent of spread in networks demarcate these networks in three broad classes of low, medium, and high effective density. We show that networks with densities above and below certain thresholds are amenable to simple heuristics or even simple random choices, whereas in between the two extremes is a region that require better and efficient methods for effective solutions of spread maximization.

Towards the sparse end of the effective density region we show that even the optimal methods do not result in significant extent of spread. Moreover, real world networks like social networks, communication networks, the web among others, are mostly sparse and generally have skewed degree distributions. On the one hand, the sparsity of the network inhibits significant extent of spreads for spreading models like the independent cascade even at high probabilities of influence. On the other hand, the degree skewness makes local structural measures like degree and eigenvalues perform as good as the optimal or greedy approximation[26].

In very dense networks, we find that, due to the uniformity of local structure of nodes, extent of spread by any method, including the optimal, are very similar. We find that proximity networks, like the ones used for studying animal social behavior or disease spread in human networks, are usually very dense. These networks are distinct from the electronic, communication, and other social networks mentioned above. We find that in such networks the extent of spread is uniformly high for almost any spread initiator. Hence, the optimal or its greedy approximation method result in the extent of spread very similar to the one obtained by using a random set of spread initiators in expectation.

In between the low and high effective density regions is an intermediate range of densities where the behavior of the optimal spread method is markedly different from random spread methods. Moreover, at some densities in this region the difference between the optimal and expected spread is as large as the approximation ratio. Hence, it is this intermediate range of densities for which application of sophisticated and computationally intense methods is necessary for finding good spread initiators efficiently. However, even in this intermediate range of densities we find that the particular structure of the networks make local structural measures like degree, variants of degree heuristics, and simple estimates like eigenvalues, to be good estimates of the optimal spread. Moreover, observing the relatively well defined modularity of the network structure in this intermediate region for block mixture model we get a better understanding of what makes certain heuristics more effective than others. Thus, taking advantage of the insights provided by this rigorous experimental analysis, better structural based heuristics can be devised that are efficient and easier to evaluate.

To summarize, we experimentally showed that the optimal extent of spread in sparse networks is achieved by the high degree nodes and in very dense networks it is achieved by any choice of spread initiators for a particular param-

eterization of the spreading model. Hence, for networks of these effective density ranges we do not need to use computationally challenging methods to find the best spreaders to maximize the extent of spread. This result leads to the discovery of an intermediate effective density range where the spread in networks is sensitive to the identity of the spread initiators. We further verify that the general heuristics for finding good spread initiators work very well for most networks due to their inherent modular structure. Hence, we can simplify a computationally hard problem of finding critical individuals for maximizing spread by limiting the application of computationally expensive methods to within a certain range of densities. This work also gives us the basis to further explore the effect of network structure on the extent of spread, both empirically and theoretically and to design algorithms that take advantage of this connection.

REFERENCES

[1] J. Aspnes, N. Rustagi, and J. Saia. Worm versus alert: Who wins in a battle for control of a large-scale network?

[2] A.-L. Barabási. *Linked*. Perseus Books Group, 2002.

[3] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, 2005.

[4] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[5] E. Berger. Dynamic monopolies of constant size. *Journal of Combinatorial Theory Series B*, 83:191–200, 2001.

[6] C. Budak, D. Agrawal, and A. E. Abbadi. Limiting the spread of misinformation in social networks.

[7] L. Chen and K. Carley. The impact of social networks in the propagation of computer viruses and countermeasures. *IEEE Trasactions on Systems, Man and Cybernetics*, forthcoming.

[8] R. Cohen, S. Havlin, and D. ben Avraham. Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 2003.

[9] Z. Dezsö and A.-L. Barabási. Halting viruses in scale-free networks. *Physical Review E*, 65(055103(R)), 2002. DOI: 10.1103/PhysRevE.65.055103.

[10] P. Domingos. Mining social networks for viral marketing. *IEEE Intelligent Systems*, 20:80–82, 2005.

[11] P. Domingos and M. Richardson. Mining the network value of customers. In *Seventh International Conference on Knowledge Discovery and Data Mining*, 2001.

[12] N. Eagle and A. Pentland. Reality mining: Sensing complex social systems. *Journal of Personal and Ubiquitous Computing*, 2006.

[13] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[14] S. Eubank, H. Guclu, V. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:429:180–184., Nov 2004. Supplement material.

[15] I. R. Fischhoff, S. R. Sundaresan, J. Cordingley, H. M. Larkin, M.-J. Sellier, and D. I. Rubenstein. Social relationships and reproductive state influence leadership roles in movements of plains zebra (*equus burchellii*). *Animal Behaviour*, 2006. in press.

[16] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001.

[17] J. Goldenberg, B. Libai, and E. Muller. Using complex systems analysis to advance marketing theory development. *Academy of Marketing Science Review*, 2001.

[18] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *CoRR*, abs/1006.0234, 2010.

[19] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[20] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[21] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 491–501, New York, NY, USA, 2004. ACM Press.

[22] Habiba, Y. Yu, T. Y. Berger-wolf, and J. Saia. Finding spread blockers in dynamic networks, 2010.

[23] G. Hartvigsen, J. Dresch, A. Zielinski, A. Macula, and C. Leary. Network structure, and vaccination strategy and effort interact to affect the dynamics of influenza epidemics. *Journal of Theoretical Biology*, 246(2):205 – 213, 2007.

[24] P. Holme. Efficient local strategies for vaccination and network attack. *Europhys. Lett.*, 68(6):908–914, 2004.

[25] M. O. Jackson and B. W. Rogers. Relating network structure to diffusion properties through stochastic dominance. *The B.E. Journal of Theoretical Economics*, 2007.

[26] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[27] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, pages 228–237, New York, NY, USA, 2006. ACM Press.

[28] L. Li, D. Alderson, J. C. Doyle, and W. Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2(4):431–523, 2005.

[29] N. M. and G. M. Finding and evaluating community structure in networks. *Phys. Rev.*, 69, 2004.

[30] R. M. May and A. L. Lloyd. Infection dynamics on scale-free networks. *Physical Review E*, 64(066112), 2001. DOI: 10.1103/PhysRevE.64.066112.

[31] Y. Moreno, M. Nekovee, and A. F. Pacheco. Dynamics of rumor spreading in complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 69(6):066130, 2004.

[32] M. Morris. Epidemiology and social networks:modeling structured diffusion. *Sociological Methods and Research*, 22(1):99–126, 1993.

[33] M. E. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(016128), 2002. DOI: 10.1103/PhysRevE.66.016128.

[34] K. Nowicki and T. A. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96:1077–1087, 2001.

[35] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200–3203, Apr 2001.

[36] B. A. Prakash, H. Tong, N. Valler, M. Faloutsos, and C. Faloutsos. Virus propagation on time-varying networks: Theory and immunization algorithms. In *ECML/PKDD (3)*, pages 99–114, 2010.

[37] E. M. Rogers. *Diffusion of Innovations*. Simon & Shuster, Inc., 5th edition, 2003.

[38] H. Tong, B. A. Prakash, C. E. Tsourakakis, T. Eliassi-Rad, C. Faloutsos, and D. H. Chau. On the vulnerability of large graphs. In *ICDM*, pages 1091–1096, 2010.

[39] T. W. Valente and W. Saba. Campaign recognition and interpersonal communication as factors in contraceptive use in bolivia. *Journal of Health Communicatio*, 2001.

[40] N. Valler, B. A. Prakash, H. Tong, M. Faloutsos, and C. Faloutsos. Epidemic spread in mobile ad hoc networks: Determining the tipping point. In *Networking (1)*, pages 266–280, 2011.

[41] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[42] D. H. Zanette. Dynamics of rumor propagation on small-world networks. *Phys. Rev. E*, 65(4):041908, Mar 2002.