Modeling Community Growth: densifying graphs or sparsifying subgraphs?

Róbert Pálovics[1,2] András A. Benczúr[1,3]
[1]Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI)
[2]Technical University Budapest
[3]Eötvös University Budapest
{rpalovics, benczur}@ilab.sztaki.hu

ABSTRACT

In this paper we model the properties of growing communities in social networks. Our main result is that small communities have higher edge density compared to random subgraphs and their edge number follows power law in the number of nodes. In other words, *the smaller the community, the larger the relative density.*

Our observation resembles the densification law of Leskovec, Kleinberg and Faloutsos who show that the average degree *increases* super-linearly as the size of the network grows. In our settings, however, densification is natural since the average degree of a *random* subgraph grows linearly. In contrary, sublinear growth translates to *increased relative density* in smaller subgraphs.

Our experiments are carried over Twitter retweets and hashtags as well as a detailed music consumption log from Last.fm. In addition to the social network of Twitter followers and Last.fm friends, key in our experiments is that community subgraphs are defined by media use.

We give theoretical results and simulations to explain our findings. The observed edge density can be explained by a mixture of epidemic growth that infects a uniform random neighbor of the community and a low probability selection of a completely new, isolated element. We also explore the relation of graph densification and subgraph sparsification by simulations over graphs of the Stanford Large Network Dataset Collection.

General Terms

Measurement, Theory

Keywords

Communities, Information spread, Power law, Densification law, Twitter, Last.fm, Community subgraphs, Complex networks, Social networks, SNAP

1. INTRODUCTION

1.1 Densification and sparsification

Part of the appeal of Web 2.0 is to find other people who share similar interests. As an example, Last.fm organizes its social network around music recommendation: users may automatically share their listening habits and at the same time grow their friendship. Based on the profiles shared, users may see what artists friends really listen to the most. Companies such as Last.fm use this data to organize and recommend music to people.

While there are several large network datasets available for research, only a few contain temporal information. We exploit the timely information gathered from services of Twitter and Last.fm to obtain microscopic measurements of influence propagating subgraphs of the social network. We define sequences of subgraphs by selecting users that have listened to the same artist, retweeted certain message or used a given hashtag. In this way we obtain evolving communities ordered in time in a fixed social network.

Our main result is a "subgraph sparsification law" of evolving community subgraphs. In time ordered subgraph sequences of the Twitter and Last.fm networks, we measure an increased edge density compared to the average edge density of the whole network. The edge density, i.e. the average degree of a node within the community follows power law of the node count. The exponent is less than two, hence the edge density growth is slower than quadratic and the *relative* density decreases, larger communities are relatively sparser than smaller communities. To understand the distinction, let us consider a random subgraph of the same size n as a selected community. As n approaches the size of the underlying network, the community and random subgraphs will cover roughly the same edges. For smaller n, hence the density of the community is *above* that of the random subgraph. In this sense, small communities that may only choose from a small n intra-community contacts are *relative* denser than the larger ones. Both the absolute and the relative density follow power law, since the number of edges in a random subgraph is quadratic.

We experiment over two large data sets. In case of Last.fm, our experiments are carried over the two-year "scrobble" history and friendship network of 70,000 Last.fm users with public profile. Last.fm's service is unique in that we may obtain a detailed timeline of how the fan community of an artist grows in time over the network.

Twitter, a mixture of a social network and a news media [13], has in the past years became the largest medium where users may spread information along their social contacts. In our experiments we use the data set of [1] that consists of the messages and the corresponding user network of four global events. We extend the tweet data with the list of followers of users with public profile who posted at least one message in the tweet dataset. The anonymized network with information spreading subgraphs is available at `https://dms.sztaki.hu/en/download/twitter-influence-subgraphs`.

As introduced before, in Last.fm and Twitter community subgraphs, we measure increasing edge density. As in [18], our subgraphs follow the densification law. However, the relative density *decreases* compared to the average edge density of the whole network. Unlike previous models of network growth, in our experiments the network is fixed and as certain information appears in this network, subgraphs are defined as the set of infected nodes. While the average degree is increasing as more nodes join the graph, this may happen for the simple reason that as larger part of a pre-existing network is explored, more connections are found for each node. Our explanation is similar to that of [21] where a sequence of subgraphs is observed as the network is gradually explored.

As a conclusion, the observed edge density can be explained by a mixture of epidemic growth that infects a random neighbor of the community regardless of the age of its infection and a low probability selection of a completely new, isolated element to the community. We also measure the importance of new isolated nodes and show that initially they dominate the communities.

We find an explanation of the community edge density in network models where new connections tend to close short paths. Such models are the forest fire one [18], the triangle closing variants of [15] and, if we add an edge to the prototype as well, the copying model of [12].

While our prime goal is to model the way communities build in social media, our models have surprising connections to densifying graphs [18, 8], and subgraph sampling [21].

Edges in the Last.fm data are timestamped. This gives us the possibility to investigate the original network densification law in case of Last.fm. We further investigate the relation of network densification and subgraph sparsification by epidemic simulations over graphs of the Stanford Large Network Dataset Collection and observe that simulated information spread in these networks follows the same power law edge density as seen in real communities.

Network growth can be considered as community growth in an unobservable hidden background network. For example, people join social networks (Facebook, LinkedIn, etc.) and expose their connections; organizations and companies exposed their relationship by gradually opening their websites in the past decade. Certain networks that are hard to fit into this category include scientific publications; indeed, the epidemic simulations in these graphs give somewhat less self-explaining exponents.

The rest of this paper is organized as follows. First we give a preview of our main observations, followed by the survey of related results. In Section 2 we give our new models for community growth and enumerate some theoretical consequences of different models of the underlying network. In Section 3 we describe our Last.fm and Twitter data that we use in our measurements in Section 4. The relations of the observations and models are discussed in Section 5.

1.2 Summary of main observations

1. "Densifying" community subgraphs with edge number following power law of node number. Note that actually the smaller subgraphs have higher *relative* density compared to a random subgraph of the same size. This difference however vanishes with the community growth, the subgraph "sparsifies".
2. Power law fraction of nodes with at least one edge within the community, with exponent greater than one. This means that initially a large fraction of the nodes are disconnected and these nodes quickly connect to one another.
3. The edge number in a community as the function of the number nodes with at least one edge also follows power law. Surprisingly, the exponent of this process is the same as the Leskovec-Kleinberg-Faloutsos [18] densification exponent and the exponent of an epidemic spread subgraph. In other words, information spreading over a network and the dynamic growth of the network are similar and closely related processes. The network itself can be considered as a community in a hidden social network.
4. Constant expansion: the number of edges leading out from the set of infected nodes is linear as long as the subgraph is not very large.

1.3 Related results

Bonchi [4] summarizes the data mining aspects of research on social influence. He concludes that "another extremely important factor is the temporal dimension: nevertheless the role of time in viral marketing is still largely (and surprisingly) unexplored", an aspect that is key in our result.

Newman reviews the theoretical background of power-law functions and distributions observed in empirical datasets in [20, 7].

As a social media service, Twitter is widely investigated for influence and spread of information. Twitter influence as followers has properties very different from usual social networks [13]. Deep analysis of influence in terms of retweets and mentions is given in [5]. Notion of influence similar to ours is derived in [6, 2] for Flickr and Twitter cascades, respectively. Cha et al. [5] define influence as "...the power of capacity of causing an effect in indirect intangible ways...". In their key observation, the influence of a user is best characterized by the size of the audience who retweets rather than the size of the follower network. We use the Twitter collection of [1] in our experiments.

Our results build on the measurements and theoretical explanations of network densification detailed in [18, 17, 19, 16]. First of all, these results state that graphs densify over time, i.e. the number of edges grow super-linearly while the average distance *shrinks* in evolving real world networks. In contrast to this observation, older network models assumed that evolving graphs have constant average degree and slowly *growing* diameter. They conclude that it is the degree sequence and not the edge sequence that has effect on the diameter of the graph. In [18] two probabilistic generative models are presented, the Community Guided Attachment and the Forest Fire model, that explain edge densification.

Dorogovtsev and Mendes calls edge densification the "accelerated growth" of the network [8]. They introduce theoretical relations between the exponent of the power-law degree distribution and the observed temporal edge densification exponent. Their computations are based on the simple assumption that the degree distribution of the graph is a power-law function of the size of the graph.

More empirical observations of densification laws can be found in [10, 22].

Pedarsani et al. investigates densification law in [21]. They state that edge densification laws can be caused by the fact that measurements on real networks are usually carried out on edges samples from the whole network. In other words, they believe that densification may arise as a feature of the common edges sampling procedure to measure dynamic networks. They show that network growth can be a direct consequence of the sampling process, therefore the sampling process itself is a plausible explanation of network densification laws.

Our experiments differ from all three lines of research (Leskovec et al., Dorogovtsev and Mendes, and Pedarsani et al.) in that we investigate a large number of coexisting subgraphs of a network that we may even consider fixed with only the communities evolving inside. Our communities show the “densification” as in the above results, however, similar to the observation of [21], we claim that the small graphs are in fact relative denser compared to the larger ones.

The results of Leskovec et al., in our terminology, consider extra-community edges as phantom nodes and phantom edges, a part of the network that is not covered by the dataset. While in a large network, this part has indeed a minor effect on the properties of edge densification, they play key role in our investigation of evolving communities.

2. MODELS FOR COMMUNITY GROWTH

2.1 Underlying network models

First we shortly summarize three main types of models for the purpose of community growth in an evolving network: concentrated degree, triangle closing and preferential attachment networks.

Certain network models impose constant degree, for example the small world models [23, 11]. Concentrated degree distribution arises in Erős-Rényi graphs [9].

Certain models build the graph by selecting edges that close triangles or short paths as [15]. The copying model [12] also falls in this category since

The main preferential attachment model is the Barabási-Albert one [3]. There the probability of connecting to a node is proportional to its degree, in other words edges connect to subgraphs based on their density.

2.2 Random node selection

We intend to investigate a model with a fixed underlying network. Nodes join after each other to the community. In every step we select a new joining node uniformly at random. In case of Last.fm that means uniform artist listening. In Twitter this model is equivalent to users that post tweets with a certain hashtag independently from each other. In this case, the expected value of the number of edges in the community is power law but with exponent equal to 2. This can be easily proven. Let E and N mean the total number of nodes and edges in the social network. Let user i an user j be part of the community with probability p, independently. The expected total number of nodes and edges in the subgraph is

$$\langle n\rangle = N \cdot p, \qquad \langle e\rangle = E \cdot p \cdot p,$$

therefore

$$\langle e\rangle \sim n^2.$$

It means that when we pick nodes uniform randomly, the average degree within the community is linear function of the subgraph size.

2.3 Epidemic spread

In the concentrated degree distribution models, the increase in the number of edges by epidemic spread is at least one and at most the maximum degree (or an upper bound such that higher degrees are very unlikely). Hence the number of edges in the community $e(n)$ grow linear with the size n.

To model an epidemic spread in the preferential attachment model, we use our observation that the edge expansion is constant, and hence the average degree within the community is equal to the average degree outside. For this reason, in the preferential attachment model, edges are equally likely to connect to any node and the results of the previous subsection apply. Notice that the observation works only under our assumption of constant expansion. In other models where infection may reach high degree nodes fast, we may have a higher probability for an edge connecting into the community.

Finally in the short path closing models, a new node u joining the community will connect to several of the close neighbors of a contact w. Let us select a contact w from the community A and let k denote the expected fraction of “close” contacts of w that are also shared by u. In this intuitive notion, the increase of the number of edges after u joins the community is hence

$$\Delta e(n) = k \cdot d(w, A). \tag{1}$$

If we assume that $d(w, A)$ is the average degree within A, we obtain $\Delta e(n) = k \cdot e(n)/n$, and the solution of the above equation becomes

$$e(n) = \text{const} \cdot n^k, \tag{2}$$

that is the edge density exponent is the same as the short path closing fraction k.

The value of k generalizes the clustering coefficient and must be at least as large in average. In the triangle closing model, all new edges close a triangle and hence k is equal to the clustering coefficient.

In a mixture of epidemic spread and random node selection, the edge density stays below that of epidemic spread. For concentrated degree distributions and preferential attachment graphs, the exponent remains the same one and two, respectively, with only a smaller constant in the edge count. For the short path closing models, if we follow the epidemic spread with probability c, we simply replace k by $c \cdot k$ in equation (1) and hence we may obtain exponents lower than the clustering coefficient.

3. DATA SETS

3.1 Last.fm

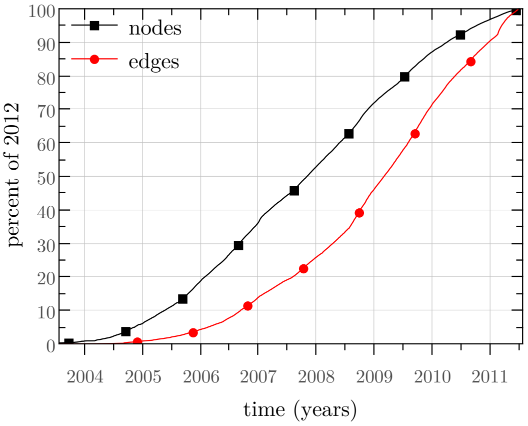


Figure 1: The number of the users and friendship edges in time as the fraction of the values at the time of the data set creation (2012) in the Last.fm dataset.

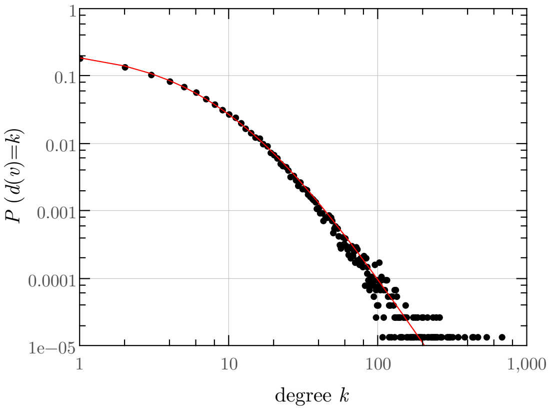


Figure 2: Degree distribution of the Last.fm social network. The distribution follows shifted power law with exponent $\alpha = 3.8$. The estimated shift is $s = 13$.

Last.fm became a relevant online service in music based social networking. The idea of Last.fm is to create a recommendation system based on plugins nearly for all kind of music listening platforms. For registered users it collects, “scrobbles”[1] what they have listened. Each user has its own statistics on listened music that is shown in her profile. Most user profiles are public, and each user of Last.fm may have friends inside the Last.fm social network. We focus on two types of user information,

- the timeline information of users: user u “scrobbled” artist a at time t (u, a, t),
- and the social network of users.

Our data set hence consists of the contacts and the musical taste of the users. For privacy considerations, throughout our research, we selected an anonymous sample of users. Anonymity is provided by selecting random users while maintaining a connected friendship network. We set the following constraints for random selection:

- User location is stated in UK;
- Age between 14 and 50, inclusive;
- Profile displays scrobbles publicly (privacy constraint);
- Daily average activity between 5 and 500.
- At least 10 friends that meet the first four conditions.

The above selection criteria were set to select a representative part of Last.fm users and as much as possible avoid users who artificially generate inflated scrobble figures. In this anonymized data set of two years of artist scrobble timeline, edges of the social network are undirected and timestamped by creation date (Fig. 1). Note that no edges are ever deleted from the network.

The number of users both in the time series and in the network is 71,000 with 285,241 edges. The average degree is therefore 8. The time series contain 979,391,001 scrobbles from 2,073,395 artists and were collected between 01 January 2010 and 31 December 2011. Note that one user can scrobble an artist at different times. The number of unique user-artist scrobbles is 57,274,158.

As the dataset is based on our selection criteria. That means it is not a simple connected part of the network, but a representative part of it. Furthermore, as the edges are timestamped, we not only see a few snapshots of the network, but have a deeper view on the process.

The degree distribution of the underlying social network follows shifted power law distribution

$$P(d(v) = k) = C \cdot (k + s)^{\alpha},$$

with exponent $\alpha = 3.8$ and shift $s = 13$. The relatively large shift is the result of our selection rules.

3.2 Twitter

The dataset was collected by Aragón et al. [1] using the Twitter API that we extended by a crawl of the user network. Our data set hence consists of two parts:

- *Tweet dataset:* tweet text and user metadata on four main global events *15O*[2], *20N*[3] *occupywallstreet*[4], *Yo Soy 132*[5].
- *Follower network:* The list of followers of users who posted at least one message in the tweet dataset.

Table 1 shows the number of users and tweets in case of each dataset. One can see that a large part of the collected tweets are retweets. Table 2 contains the size of the crawled social networks. Note that in all four networks, the average in- and outdegree is relatively high. Fig. 4 shows the in-

[1] The name “scrobbling” is a word by Last.fm, meaning the collection of information about user listening.

[2] http://en.wikipedia.org/wiki/15_October_2011_global_protests

[3] http://en.wikipedia.org/wiki/20-N

[4] http://en.wikipedia.org/wiki/Occupy_Wall_Street

[5] http://en.wikipedia.org/wiki/Yo_Soy_132

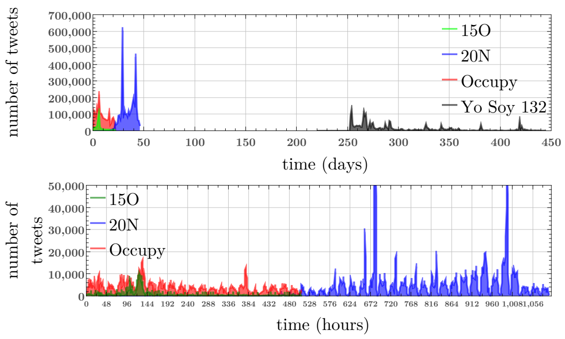


Figure 3: Temporal density of tweeting activity in the four different Twitter datasets.

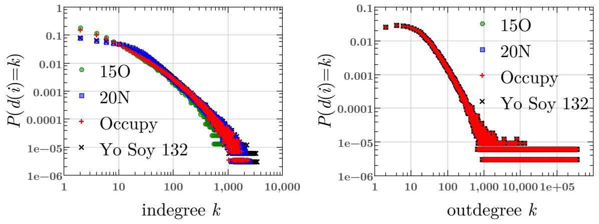


Figure 4: Degree distributions of the Twitter follower networks.

and outdegree distribution of the collected networks. Fig. 3 shows the temporal density of tweeting activity in case of the four different datasets. For each tweet, our data contains

- tweet and user ID,
- timestamp of creation,
- hashtags used in the tweet.

In case of a retweet, we have all these information not only on the actual tweet, but also on the original tweet that had been retweeted.

	15	oc	yo	20
# users	96,935	371,401	395,988	366,155
# tweets	410,482	1,947,234	2,439,109	1,947,234
# hashtags	28,014	93,706	62,008	123,925

Table 1: Sizes of the tweet time series.

	15	oc	yo	20
# users	83,640	330,677	363,452	336,892
# edges	3,093,966	16,585,837	22,054,165	18,809,308
avgdeg.	37	50	61	56

Table 2: Sizes of the follower networks.

3.3 SNAP graphs

We use the following graphs of the Stanford Large Network Dataset Collection[14]:

- **ArXiv** HepPh: Arxiv High Energy Physics paper citation network (phenomenology),
- **ArXiv HepTh**: Arxiv High Energy Physics paper citation network (theory),

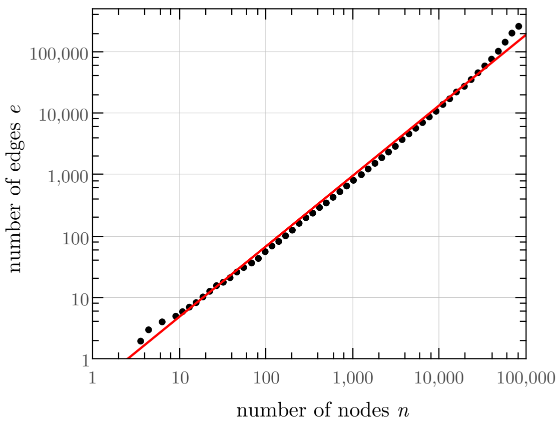


Figure 5: Network densification law in the Last.fm dataset. The number of edges is power law function of the number of nodes in the evolving social network with exponent $\beta = 1.14 - 1.17$.

- **DBLP**: DBLP collaboration network,
- **LiveJournal**: LiveJournal online social network,
- **CAIDA**: The CAIDA AS Relationships Dataset,
- **Google**: Web graph from Google,
- **EU email**: Email network from a EU research institution.

4. EXPERIMENTS

4.1 Network densification

As observed in [18], one common property of complex networks is the edge densification law. As new nodes join in, the number of edges follows a power law of the number of nodes. For Last.fm, we sort the edges by their creation time and then sort the nodes based on this list. Node by node we measure the increase of the number of edges Figure 5. Densification law holds in case of Last.fm with exponent $\beta = 1.14 - 1.17$. Note that regarding to Section 3 no edges were ever deleted from the Last.fm network. Notice that we do not have temporal information on the Twitter follower graph.

4.2 Topical communities

Next we introduce three special community related subsets and define topical communities in Last.fm and Twitter. Let $A(t)$ mean the subset of users in a social network that have adopted a certain topic before time t. As shown in Figure 6, we call a *community subgraph* the graph of users in A.

The “non-zero” component is the subgraph of users that have at least one edge within the community. This component contains all the edges within the community. A contains only isolated nodes besides the “non-zero” component.

The “main component” of A is measured as the one reachable through directed influence edges from the first infected node. With high probability, this is also the largest component. Note that we consider directed reachability, i.e. we do not merge two initial seeds of infected nodes into the same component when they both reach the same new node. Later we investigate the properties of the community subgraph,

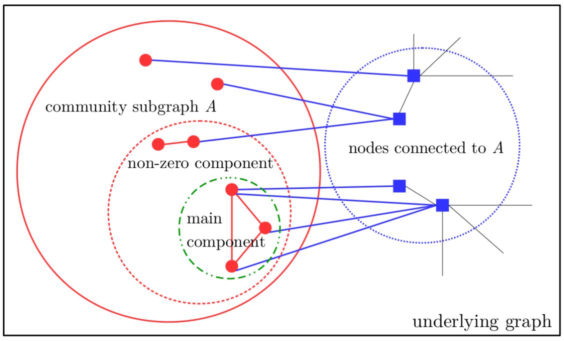


Figure 6: Important subsets of a community subgraph.

the non-zero component, and the main component.

In Last.fm, commmunities are formed by users that have listened to the same artist. $A(t)$ is the subset of users that have scrobbled a given artist at least once before time t.

In case of Twitter a community subgraph is formed by users that have tweeted a given hashtag before time t. In other words we investigate artist subgraphs in Last.fm, and hashtag subgraphs in the Twitter follower network.

In what follows we introduce measurements that result power-law functions related to community subgraphs. Table 3 summarizes the notations and our results in the Last.fm dataset. Table 4 shows the measured exponents for the four Twitter datasets. Next we introduce and investigate these power-law exponents in details. Note that as we have more hashtags than artists, our measurements are more accurate in case of Last.fm than in case of Twitter-. In Table 4 the error of the exponents are roughly 0.05.

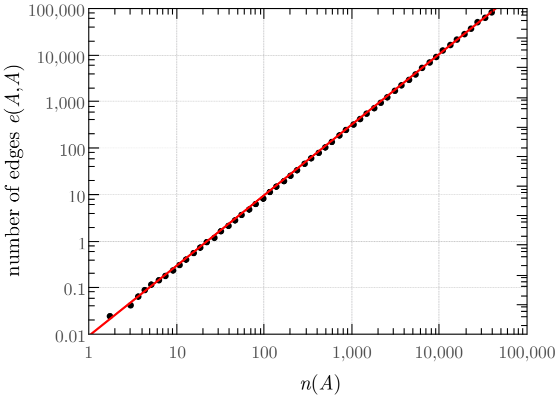
4.3 Community subgraph density

To deeper understand the properties of a community subgraph, we set up the following measurement. For each time t a new user adopts the community's topic, we measure the number of edges $e(A,A)$ in the subgraph as the function of the number of users $n = |A|$ in the subgraph. We compute function $e(n)$ for each artist in case of Last.fm and for each different hashtag in Twitter. We average the $e(n)$ curves in case of both social networks. Note that the Twitter follower graph is directed. In that case an edge is part of the subgraph if its source joined earlier to the community than its target. Figure 7 shows our results. In case of Twitter we have four different curves corresponding to the four different datasets. One of our key results is that number of edges is power-law function of the size of the community subgraph

$$e(n) \sim n^{\gamma}. \tag{3}$$

The exponent is 1.52 in Last.fm, and $1.42 - 1.5$ in Twitter communities. Note that we not only averaged the final community subgraphs, but averaged all temporal states of all community subgraphs. Our conclusion is that subgraphs of users with the same activity in a social network show power law growth. Both the number of edges and the average degree are increasing power law function of the number of nodes in the graph.

Figure 8 shows the average degree to the community of the joining node as the function of the community's size. The curves are roughly the derivative of the ones in Figure 7.



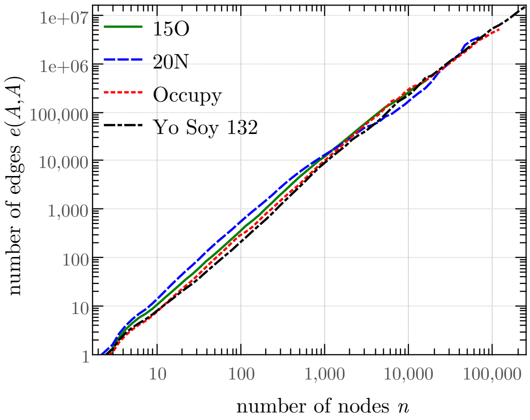


Figure 7: Community subgraph densification in the Last.fm (top) and Twitter (bottom) datasets. The number of edges is power law function of the number of nodes in a community subgraph.Top: Last.fm, Bottom: Twitter.

4.4 Non-zero degree component

We introduce another power-law result as an explanation of subgraph densification. We can measure the size of the non-zero component z as the function of the size of the subgraph n. That is the number of nodes with non-zero degrees in the subgraph. Figure 9 shows our results. $z(n)$ is a power-law function,

$$z \sim n^{\delta}. \tag{4}$$

Exponent δ is between $1.36 - 1.38$ for Last.fm artists, and 1.1 for Twitter hashtags. Equations (3) and (4) predict that edges in the non-zero component densify with another exponent β_z,

$$e(z) \sim z^{\beta_z}, \qquad \beta_z = \gamma/\delta. \tag{5}$$

We can either compute $\beta_z = \gamma/\delta$ or plot e as the function of z (see Fig. 10). In Last.fm β_z is between 1.15 - 1.17, while it is between 1.31 - 1.38 for Twitter hashtags.

4.5 Main epidemic component

As introduced in Section 4.2, the main component is measured as the one reachable through directed influence edges from the first infected node. Our next measurement (see

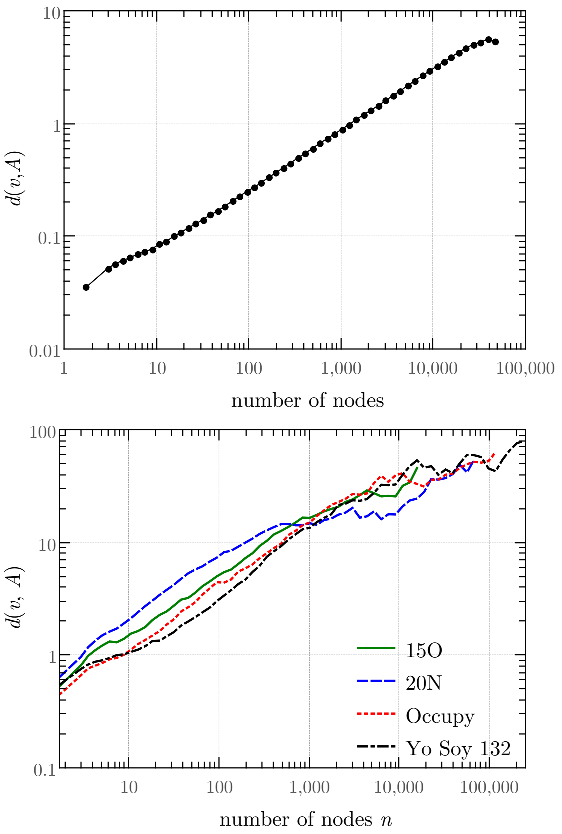


Figure 8: Average degree of the connecting node to the subgraph as the function of the community subgraph size.Top: Last.fm, Bottom: Twitter.

Fig.10) is that the number of edges ϵ in the main component is power-law function of the size its size m,

$$\epsilon \sim m^{\beta_m}. \tag{6}$$

The corresponding exponent is 1.15 for Last.fm. It is between 1.32 - 1.37 for Twitter networks.

degree distribution α	3.8
network densification β	1.14 - 1.17
subgraph densification γ	1.52
uncorrelated model γ_0	2
non-zero nodes δ	1.36 - 1.38
non-zero component densification β_z	1.15 - 1.17
main component densification β_m	1.15
epidemic β_e	1.14 - 1.15

Table 3: Summary of the most important exponents in th Last.fm dataset.

	γ	δ	β_z	β_m	β_e	β_r
Occupy	1.47	1.1	1.37	1.36	1.35	1.19 - 1.22
Yo Soy 132	1.49	1.1	1.36	1.37	1.27	1.19 - 1.25
20N	1.42	1.1	1.31	1.32	1.27	1.1 - 1.25
15O	1.5	1.07	1.38	1.37	1.32	1.16 - 1.3

Table 4: Exponents in the four Twitter datasets.

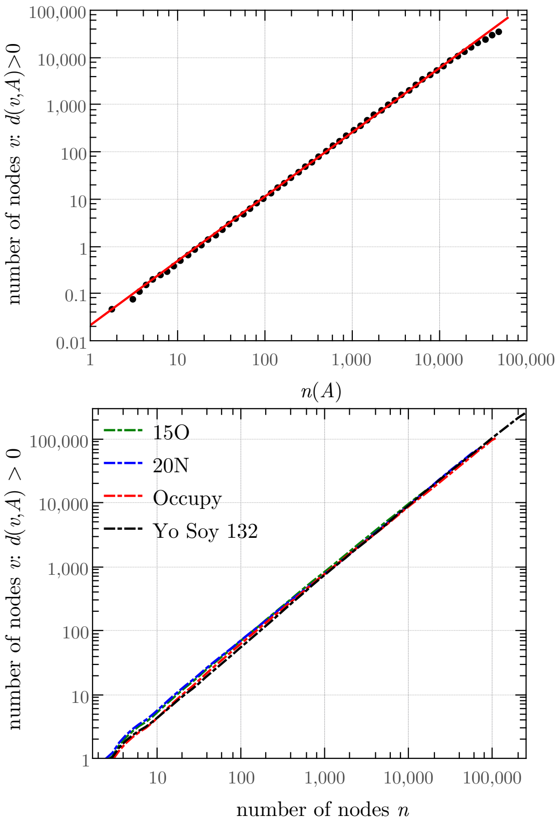


Figure 9: Number of nodes with non-zero degrees as the function of the number of nodes in a community subgraph. Top: Last.fm, Bottom: Twitter.

4.6 Constant expansion

Figure 11 shows the number of edges leading out from the Last.fm and Twitter communities as the function of the subgraph size. One can observe that in both cases the function is linear as long as long as the subgraph is not very large.

4.7 Epidemic simulations

To investigate the model introduced in Section 2, we simulated epidemic processes in Last.fm, Twitter, and SNAP networks. Starting from a uniform randomly picked node we generated infection processes. At each step we select uniform randomly a node that is not joined to the community, but connected to it in the network (see Fig. 6). Subgraph densification holds for these communities with exponent β_e. Figure 10 shows our results for Last.fm and Twitter networks. Exponents can be found in Table 3 and Table 4. Figure 12 shows our results for SNAP datasets. Table 5 summarizes the exponents for SNAP data. Figure 13 shows for each network the relation of exponent β_e and the average clustering coefficient of the network. Figure 14 shows the number of edges leading out as the function of the epidemic generated community's size.

5. DISCUSSION

In this section we discuss how the network and community densification laws relate to one another and the predictions

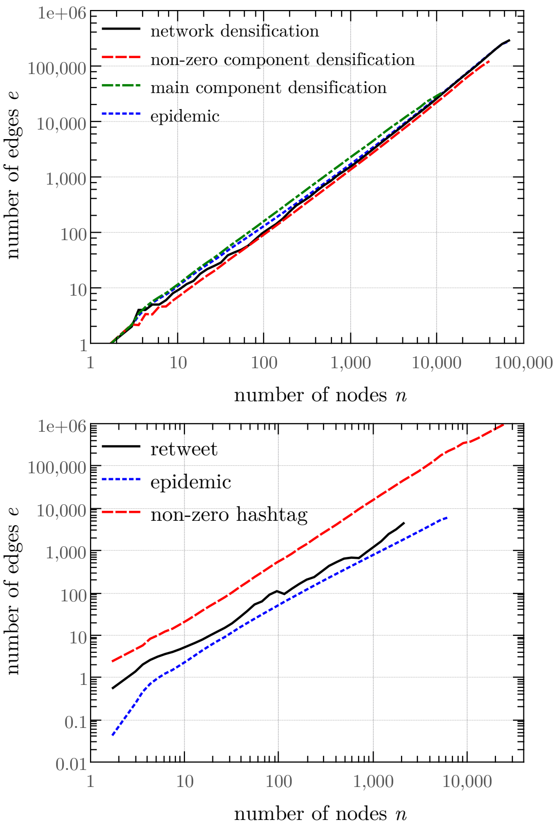


Figure 10: Comparison of different processes with similar exponents. Top: Last.fm, Bottom: Twitter.

network	clustering coefficient	β_e
Last.fm	0.18	1.14
ArXiv HepTh	0.323	1.25
ArXiv HepPh	0.283	1.2
DBLP	0.63	1.06
CAIDA	0.208	1.1
LiveJournal	0.283	1.1
Google	0.5143	1.02
Twitter Occupy	0.12	1.35
EU email	0.0671	1.06

Table 5: β_e and the clustering coefficient in case of different real-world networks.

of the model in Section 2. Tables 3-4 summarize all power law exponents that we discussed. Here we intend to focus on β, γ and δ.

Figure 16 shows in one plot the result of the epidemic model, the uniform model, and the measured artist subgraph densification law in Last.fm. As introduced in Sections 1-2, the measured curve is between the epidemic model and the uniform random model. This indicates that artist densification in Last.fm is the mixture of an epidemic and a random process. This figure also shows how the relative densification to the random model disappears from the community subgraphs. Larger artist subgraphs are relatively sparser then smaller subgraphs.

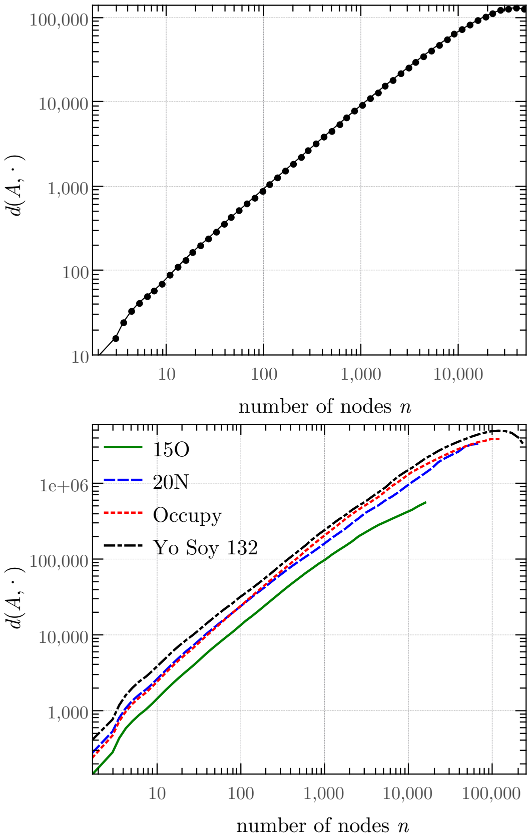


Figure 11: Number of edges leading out from the community subgraph A as the function of the subgraph size. Top: Last.fm, Bottom: Twitter.

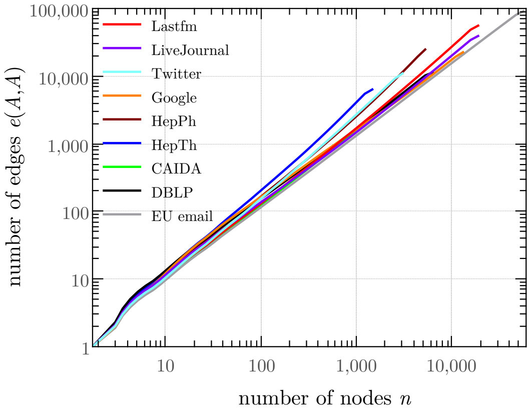


Figure 12: Results of epidemic simulations on various real-world graphs.

Next we compare the values of γ and β_e in case of Last.fm to the exponents measured in case of Twitter. As hashtags can spread with retweets, γ is closer to β_e in hashtag subgraphs than artist subgraphs. In other words information spreading is much stronger in hashtag defined communities then in artist defined ones.

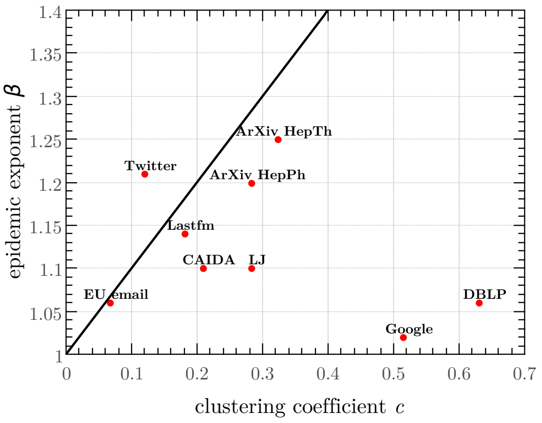


Figure 13: Relation of the epidemic exponent β an the clustering coefficient in case of the 9 different real-world networks.

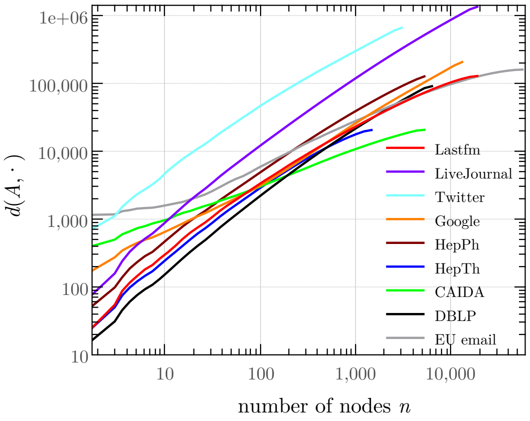


Figure 14: Number of edges leading out from the community subgraph A as the function of the subgraph size.

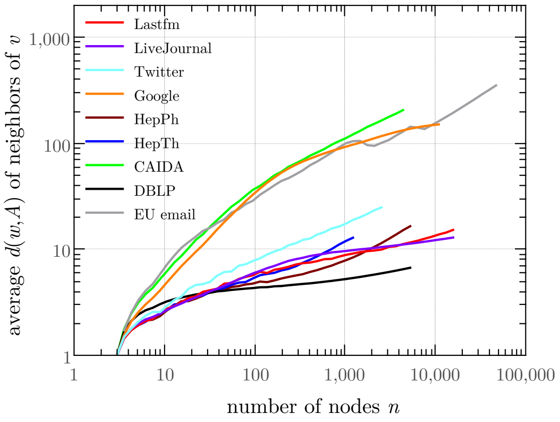


Figure 15: Average degree $d(w, A)$ of neighbors of the recently joined node.

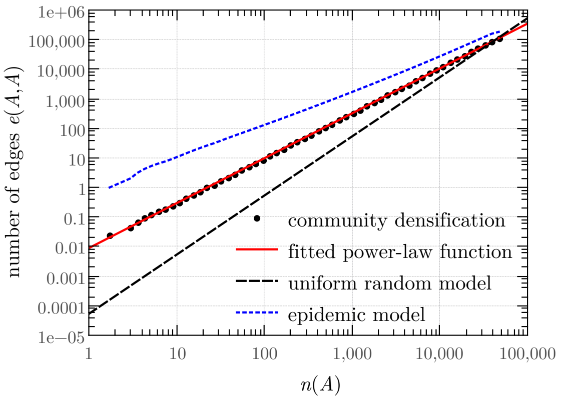


Figure 16: Difference between the measurement and the uniform model in the Last.fm dataset.

Our next observation is that regarding to Tables 3-4,

$$\beta = \beta_e = \beta_z = \beta_m, \qquad (7)$$

the four exponents are identical for Last.fm, and similar for Twitter. This result can be seen in Figure 10, where we plotted the curves corresponding to these exponents. Surprisingly, in case of Last.fm, not only the exponents, but the curves are identical. Note that in case of Last.fm we can also observe the network densification law with exponent β. Our results show that epidemic processes over the network are similar to the temporal evolution of the network. Moreover, because of (5) and (7),

$$\gamma = \beta \cdot \delta. \qquad (8)$$

This relation between the exponents means that network densification exponent β, and the non-zero exponent δ controls the subgraph densification exponent. Edge density in community subgraphs can be explained by a mixture of epidemic growth that infects a uniform random neighbor of the community and a low probability selection of a completely new, isolated element.

In case of Twitter we computed retweet community subgraphs. As shown in Figure 10, the curve of the epidemic model and the retweet subgraph densification are similar. Note that in contrast to Last.fm, the hashtag curve is over the epidemic curve. However we believe this observation is caused by the quality of the Twitter data. As it is constructed from multiple crawls, we do not have all the edges of the follower network. Moreover, we have less hashtags and retweets than artists in Last.fm.

Next we relate the densification coefficients to the clustering coefficient as in the model of Section 2. As seen in Table 5, for certain networks including Last.fm and the EU email, the two values are very close and for Twitter, even $\beta > k$, indicating a strong tendency to close short paths and connect inside a small community.

For a large number of data sets, however, the clustering coefficient is larger than the densification exponent. As we have no easy-to-define communities in these graphs, the measurements simply indicate that epidemic growth in these networks follow a somewhat different pattern.

In order to investigate graphs with $\beta < k$ further, we identify the reason for the deviation from equation (1) in

Section 2. There we assumed that the degree $d(w, A)$ of the existing member of the community who joins the new member u does not deviate from the average. In particular, $d(w, A)$ should follow the power law $e(n)/n$. In Figure 15 we see that the more a network deviates from the $\beta \approx k$ rule, the quicker a decay in the increase of the average degree happens. The effect of the decayed growth of $d(w, A)$ is lower edge count compared to our model. While the behavior of epidemic spread in these networks is not directly in the scope of this paper, we emphasize this finding as a potential phenomenon that needs further explanation.

6. CONCLUSIONS

In this paper we investigated the properties of growing communities in social networks. We used data from popular social networking sites Last.fm and Twitter to study in details the evolution of communities in large graphs.

We introduced the community subgraph sparsification law. To understand this effect, we carried over numerous of measurements, that resulted various power-law functions between specific quantities related to community subgraphs. We explained the theoretical background and the relation of these power-law exponents. The results of our experiments show that the observed edge density in a community can be explained by a mixture of epidemic growth that infects a uniform random neighbor of the community and a low probability selection of a completely new, isolated element. According to our results epidemic driven community growth is similar to the original network densification: network growth can be considered as community growth in an unobservable social network.

7. ACKNOWLEDGEMENTS

To the Last.fm team for preparing us this volume of the anonymized data set that cannot be efficiently fetched through the public Last.fm API.

To Andreas Kaltenbrunner for providing us with the Twitter data set [1].

Research supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No 288956) and by the grant OTKA NK 105645.
Support from the "Momentum - Big Data" grant of the Hungarian Academy of Sciences.
The work of Robert Palovics reported in this paper has been developed in the framework of the project "Talent care and cultivation in the scientific workshops of BME" project. This project is supported by the grant TAMOP - 4.2.2.B-10/1–2010-0009. Work conducted at the Eötvös University, Budapest was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/-KONV-2012-0013). The research was carried out as part of the EITKIC_12-1-2012-0001 project, which is supported by the Hungarian Government, managed by the National Development Agency, financed by the Research and Technology Innovation Fund and was performed in cooperation with the EIT ICT Labs Budapest Associate Partner Group. (`www.ictlabs.elte.hu`)

8. REFERENCES

[1] P. Aragón, K. E. Kappler, A. Kaltenbrunner, D. Laniado, and Y. Volkovich. Communication dynamics in twitter during political campaigns: The case of the 2011 spanish national election. *Policy & Internet*, 5(2):183–206, 2013.

[2] E. Bakshy, J. M. H., W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

[3] A.-L. Barabási, R. Albert, and H. Jeon. Mean-field theory for scale-free random network. *Physica A*, 272:173–187, 1999.

[4] F. Bonchi. Influence propagation in social networks: A data mining perspective. *IEEE Intelligent Informatics Bulletin*, 12(1):8–16, 2011.

[5] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[6] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing social cascades in flickr. In *Proceedings of the first workshop on Online social networks*, pages 13–18. ACM, 2008.

[7] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[8] S. Dorogovtsev and J. Mendes. Accelerated growth of networks in handbook of graphs and networks: From the genome to the internet., 2002.

[9] P. Erdős and A. Rényi. On the evolution of random graph. *Math. Inst.*, 1960.

[10] J. S. Katz. Scale-independent bibliometric indicators. *Measurement: Interdisciplinary Research and Perspectives*, 3(1):24–28, 2005.

[11] J. Kleinberg. The Small-World Phenomenon: An Algorithmic Perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.

[12] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1–10, 2000.

[13] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[14] J. Leskovec. Stanford large network dataset collection. *URL http://snap. stanford. edu/data/index. html*, 2011.

[15] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470. ACM, 2008.

[16] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *The Journal of Machine Learning Research*, 11:985–1042, 2010.

[17] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM, 2006.

[18] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

[19] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.

[20] M. E. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[21] P. Pedarsani, D. R. Figueiredo, and M. Grossglauser. Densification arising from sampling fixed graphs. *ACM SIGMETRICS Performance Evaluation Review*, 36(1):205–216, 2008.

[22] S. Redner. Citation statistics from more than a century of physical review. *arXiv preprint physics/0407137*, 2004.

[23] D. J. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.