Densification Arising from Sampling Fixed Graphs

Pedram Pedarsani
School of Computer and Communication Sciences
EPFL
Lausanne, Switzerland
pedram.pedarsani@epfl.ch

Daniel R. Figueiredo
Fed. Univ. of Rio de Janeiro
COPPE/PESC
Rio de Janeiro, Brazil
daniel@land.ufrj.br

Matthias Grossglauser
Internet Laboratory
Nokia Research Center
Helsinki, Finland
matthias.grossglauser@nokia.com

ABSTRACT

During the past decade, a number of different studies have identified several peculiar properties of networks that arise from a diverse universe, ranging from social to computer networks. A recently observed feature is known as network *densification*, which occurs when the number of edges grows much faster than the number of nodes, as the network evolves over time. This surprising phenomenon has been empirically validated in a variety of networks that emerge in the real world and mathematical models have been recently proposed to explain it. Leveraging on how real data is usually gathered and used, we propose a new model called *Edge Sampling* to explain how densification can arise. Our model is innovative, as we consider a fixed underlying graph and a process that discovers this graph by probabilistically sampling its edges. We show that this model possesses several interesting features, in particular, that edges and nodes discovered can exhibit densification. Moreover, when the node degree of the fixed underlying graph follows a heavy-tailed distribution, we show that the Edge Sampling model can yield power law densification, establishing an approximate relationship between the degree exponent and the densification exponent. The theoretical findings are supported by numerical evaluations of the model. Finally, we apply our model to real network data to evaluate its performance on capturing the previously observed densification. Our results indicate that edge sampling is indeed a plausible alternative explanation for the densification phenomenon that has been recently observed.

Categories and Subject Descriptors

H.1 [**Models and Principles**]: General

General Terms

Experimentation, Measurement

Keywords

network modeling, densification, edge sampling

SIGMETRICS'08, June 2–6, 2008, Annapolis, Maryland, USA.

1. INTRODUCTION

Over the past decade, investigations in different fields have focused on studying and understanding networks that arise in their respective domains, ranging from biological to social to technological networks. Surprisingly at first, many of these networks exhibit common topological features, such as a heavy tailed degree distribution and the small world effect[1]. Surveys in this area, known as *Complex Networks*, include the work of Albert and Barabasi [1], Newman [10] and Boccaletti [3].

Most of the work on such complex networks has focused on static scenarios, where a single snapshot of the network is considered for investigation. More recently, the focus has changed to consider dynamic scenarios, where the network evolves with time. Indeed, most real networks evolve over time, as edges and nodes can be added or deleted from the system. In this context, a recently observed feature is *densification*, which occurs when the number of edges grows faster than the number of nodes. More precisely, the average degree of the network grows with time. This surprising phenomenon has been empirically validated by the recent work of Leskovec et al. [9], where densification is observed in six large datasets. Even more surprising is the fact that the observed densification exhibited a very specific and precise relationship, namely, a power law of the form $e(t) \sim n(t)^{\alpha}$, where $e(t)$ and $n(t)$ denote the number of edges and nodes of the network at time t and α is a constant greater than 1. We refer to power-law densification with exponent α as α-densification.

Besides investigating and discovering peculiar features of graphs that arise in real data, another direction of work is the development of plausible mathematical models that can explain or capture the observed phenomena. Several models have been proposed to capture different features, such as the preferential attachment model [5, 2]. With respect to network densification, Leskovec et al. [9] have proposed mathematical models for network growth that result in α-densification. A key element of their model is the fact that newly created nodes establish more edges than nodes that were created earlier. They show that this model can lead to a power law densification over time.

Densification is a surprising feature. It implies that the average node degree grows without bound, which is counter intuitive in many real settings. For example, does it make sense that an individual's number of social ties depends on the size of the total population of the planet? Even more counter intuitive is the notion that densification can lead to a decrease of the network diameter (or of the average distance) as the network grows. Does it make sense that the average distance between Internet domains decreases as the Internet grows? Despite these intuitive doubts, the evidence from

[1]A network is considered a small world if it exhibits a high clustering coefficient and short pair-wise distances.

several disparate datasets is convincing and solid. We feel therefore that reconciling the observed phenomenon with domain intuition is a promising area of research.

Our contribution is to provide a novel explanation for the observed densification in real networks. Our explanation posits that densification may arise as a feature of a common procedure to observe - or measure - dynamic networks, rather than as a feature of the network itself. To sharpen this explanation, we show in this paper that densification can actually arise even when observing a *fixed* network that is gradually discovered through a sampling process. This sampling process captures the usual way of observing dynamic networks.

More specifically, two observations are at the heart of our approach, which explains densification through the sampling of a fixed underlying graph. The first observation concerns the way we measure networks. In fact, we argue that in many empirical studies of complex networks, it is the links (or edges) that are observed directly, and the nodes (or vertices) are *only revealed indirectly* through the observation of links. For example, most studies of email networks are based on a log of email messages. An email message exchanged between two email addresses a and b is taken as evidence of a social link (a, b). At the same time, this message *reveals* the nodes a and b if these nodes were not already known. We believe that the direct observation of edges, which at the same time gradually reveals the nodes, is a feature of most empirical studies of complex networks. We argue in this paper that this way of observing - or sampling - a network can give rise to perceived densification.

The second observation concerns the way we explain network growth. Many studies of network evolution have assumed that new nodes are added to the network in sequence, with some rules to establish links to existing nodes. However, we argue that network growth may be at least partially explained by the gradual observation of nodes and links that exist permanently "in the background". In other words, there exists a fixed underlying network that is not directly observable. For example, this network may represent the people in a large organization, and the social and professional ties that bind them. An edge of this network can be observed only once this edge "fires", e.g., a message is sent over this edge. We believe that in many situations, it is reasonable to assume that such a hidden network exists, which changes on a time-scale much longer than the sampling process. The network growth is then a direct consequence of the sampling process, i.e., the gradual discovery of this underlying network, rather than a property of the network itself.

We believe that these two features are quite universal in the study of complex networks. The goal of our paper is to shed light on the possibility that densification can at least partially be explained by the observation of a fixed network that is gradually discovered by sampling its edges, e.g., the exchange of email messages to reveal social ties.

To support this claim, we propose a model formed by a fixed underlying graph and by a process that discovers this graph by sampling its edges over time. We refer to this model as the *edge sampling model*. We consider two variants of this model, which capture two common procedures for the observation of real networks. The first variant is the *accumulation model*, where we assume that the observed network is the result of all the edges discovered since the start of the observation. As time evolves, the observed network grows and "converges" to the hidden full network. The second variant is the *modulation model*, where we assume that independent snapshots are obtained at different times, which we can view as samples of the hidden full network. We study these two variants to reflect different measurement methodologies used in studies of network evolution, and we show that both variants can lead to densification in the observed network. How the sampled networks densify depends both on the structural properties of the underlying network and on the sampling process itself.

Using a simplified instantiation of the edge sampling model, we establish analytical results indicating that the number of nodes and edges discovered indeed densify over time. In particular, we prove that densification is present for all time instances greater than a threshold. We also prove that densification converges to a power law as time increases, with an exponent that is inversely proportional to the probability mass of degree one nodes. Finally, we apply our model to real network data and show that it can fairly capture the densification observed in practice. Our results indicate that edge sampling is a plausible alternative explanation for the network densification.

As stated above, the key motivation for our model lies in the way many real networks are observed: edges of an underlying graph are observed directly, with nodes being observed indirectly. We do not claim that all networks densify because of edge sampling. Neither do we claim that all real networks densify. However, we claim that edge sampling can be a plausible explanation for the observed densification of some networks, where the observation of edges leads to the discovery of nodes from the underlying graph. In other words, we explain densification as a feature of the statistical estimation instead of a feature of the real network. In [9] Leskovec et al. propose network growth as a plausible model to explain densification.

We close this section by commenting on the difference between degree power laws, a widely researched feature of many real networks, and power-law densification. It is important to note that the two are orthogonal, in the sense that each can exist independently. For example, a sequence of d_n-regular graphs $G(n, d_n)$ with $d_n = n^{\alpha-1}$ ($\alpha > 1$) does not exhibit a degree power law, as every snapshot has constant degree over the ensemble of nodes; however, it does exhibit α-densification. On the other hand, a sequence of random graphs whose degree distribution is drawn from a distribution D with power-law tail (and with $E[D] < \infty$) does exhibit degree power law, but not densification.

Degree power laws have often been linked to the "rich get richer" phenomenon. This refers to the fact that rich (e.g., large degree) nodes tend to become richer (i.e., even larger degrees) as the network grows in size, giving rise to power law degree distributions. This is the main idea, for example, behind the preferential attachment models [5, 2], used for explaining how a single snapshot of the network is formed. This degree distribution is usually fixed, despite models that allow the graph to grow over time. However, power law densification refers to the relationship between number of nodes and edges as the network evolves (i.e., in different snapshots of the network). In other words, power-law degree distribution is the property of a single snapshot of the network, while power-law densification refers to the relationship between number of edges and nodes for a sequence of networks (e.g., because the network is growing). Notice that in classical graph models the average degree is assumed to be constant over time (i.e., the number of edges grows linearly in the number of nodes), thus no densification is present (though featuring power law degree distribution).

The remainder of this paper is organized as follows. Section 2 shows a concrete example of network densification and the problem statement. Section 3 presents the Edge Sampling model, its theoretical properties and numerical evaluations. Sections 4 presents an evaluation of the proposed model when applied to real network data. Section 5 briefly discusses the related work. Finally, Section 6 concludes the paper.

2. MOTIVATION

In order to motivate and illustrate the concept of densification, we present an example of a real network that densifies over time. In particular, we consider what is known as an *email network*. In this directed graph, nodes represent email addresses and edges represent message exchanges. The network evolves in time through the observation of new messages that are exchanged, as described below.

Let $m_{ij}(t)$ denote a message sent from email address i to email address j at time t. Moreover, let $M[t_1, t_2]$ denote the set of all messages sent in the interval $[t_1, t_2]$.
Let $G[t_1, t_2] = (V[t_1, t_2], E[t_1, t_2])$ be a directed graph defined as follows. Let $E[t_1, t_2] = \bigcup_{m \in M[t_1,t_2]} e(m)$, where $e(m_{ij}(t)) = \{(i, j)\}$. Thus, $E[t_1, t_2]$ denotes the set of directed edges that appear in the graph in the interval $[t_1, t_2]$. Similarly, let $V[t_1, t_2] = \bigcup_{m \in M[t_1,t_2]} v(m)$, where $v(m_{ij}(t)) = \{i, j\}$. Thus, $V[t_1, t_2]$ denotes the set of nodes that appear in the graph in the interval $[t_1, t_2]$.

We consider a dataset of email messages collected at the mail server of EPFL for a period of 89 weeks. The dataset is aggregated by week, such that timestamps are in the timescale of weeks (i.e., all messages sent in a particular week have the same timestamp). To filter out bogus messages (spam, mailing lists, wrong addresses, etc), only emails to or from registered EPFL personnel were considered, as well as only email addresses that both sent and received a message at least once in the observed interval of time. The filtered dataset contains a total of $23,679,417$ email messages.

Following the procedure describe above, we construct the email network by growing it one week at a time, computing $G[1, t]$, where $t = 1, \ldots, 89$. Thus, for each week t, we have a value for the number of edges $e(t) = |E[1, t]|$ and the number of nodes $n(t) = |V[1, t]|$ in the graph. Figure 1 plots the number of nodes versus the number of edges for all values of t in log-log scale. The result clearly shows a densification of the network, as the number of edges grows must faster than the number of nodes. In particular, the number of nodes grows from $n(1) = 45,782$ to $n(89) = 431,200$, while the number of edges grows from $e(1) = 115,898$ to $e(89) = 3,945,937$. Moreover, we can observe a fairly precise linear relationship between $\log e(t)$ and $\log n(t)$, which indicates that the two variables are related through a power law of the type $e(t) \sim n(t)^{\alpha}$. Let $\hat{\alpha}$ be the slope of the line obtained by performing a linear regression of the points in the plot. Thus, for the data shown in Figure 1 we have that $\hat{\alpha} = 1.57$.

As stated earlier, Leskovec et al. [9] have also observed power law relationships between the number of edges and nodes for various real networks, including an email network. The fact that power laws seem to be ubiquitous relationship between the number edges and nodes of an evolving network motivates us to search for alternative explanations for densification.

It is important to notice how the email network was grown in the example above. Indeed, the network is grown one message at a time, revealing edges at each step, while nodes are discovered through these edges. We believe that many real networks are also grown in this same way, that is, through the observation of edges. For example, in the World-Wide Web network, web documents (nodes) are discovered through the observation of hyperlinks in the documents (edges). In the Internet AS-level graph, ASes (nodes) are inferred through the observation BGP announcements (edges). In the IMDB actors to movies networks, actors and movies (nodes) are discovered when an actor plays in a movie (edges). Thus, the process of growing the network through the observation of edges (or *edge sampling*) seems fairly universal and is the key motivation for the model we introduce next.

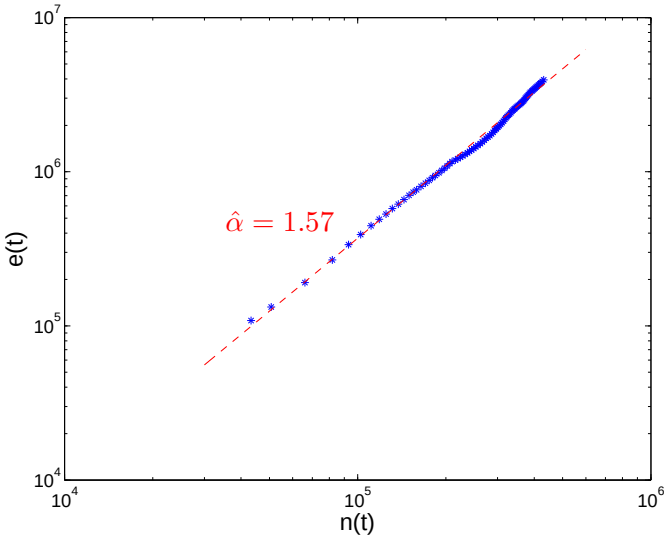

Figure 1: Evolution of the number of nodes $n(t)$ versus the number of edges $e(t)$ over time for EPFL email network.

3. THE *EDGE SAMPLING* MODEL

In this section, we propose a model for the discovery of an unknown but fixed underlying graph $G = (V, E)$. The key feature of this model is that edges are observed directly, while vertices are only implicitly revealed through its adjacent edges. This is motivated by the fact that many datasets of real networks are constructed through the observation of its edges. For this reason, we call our model *edge sampling*, since edges are directly revealed during the discovery process. The key idea is that edges of the unknown underlying graph randomly *fire*, revealing themselves to the discovery process. When an edge is sampled, the two adjacent vertices of that edge are also revealed.

We elaborate the model by considering specific properties of the underlying graph and specific properties for the sampling of edges. In particular, let $G = (V, E)$ be the underlying graph with $n_v = |V|$ denoting the total number of vertices and $n_e = |E|$ denoting the total number of edges in the graph. Let $f_D(k)$ denote the empirical degree distribution of G. We assume that all edges $e \in E$ are associated with a sampling probability p_e. We obtain a single sample $G' = (V', E')$ of the graph by including every edge $e \in E$ independently with probability p_e, and by letting V' be the set of adjacent vertices of all edges in E'. Note that p_e is the sole parameter of the proposed model and will be referred to as the *sampling parameter*.

Thus, we can define the set $E' \subset E$ to denote the set of edges that have fired in the realization when the sampling parameter is p_e, and hence discovered. Similarly, define the set $V' = \bigcup_{e \in E'} \gamma(e) \subset V$ to denote the set of vertices that have been discovered, where $\gamma(e)$ is a function that returns the set of vertices adjacent to the edge e, that is, $\gamma((u, v)) = \{u, v\}$. Thus, we can define $e(p_e) = |E'|$ and $n(p_e) = |V'|$ as the number of edges and vertices, respectively, that are discovered when the sampling parameter is p_e. In the following, we determine both $E[e(p_e)]$ and $E[n(p_e)]$, that is, the expectation of these random variables.

Since all edges are discovered with same probability p_e, we can write:

$$E[e(p_e)] = n_e p_e \tag{1}$$

However, let $E[D]$ denote the expected degree of the underlying

graph. It can be shown that $n_e = n_v E[D]/2$. Thus, we can rewrite equation (1) as follows:

$$E[e(p_e)] = n_v E[D] p_e/2 \tag{2}$$

Let p_v denote the probability that a node is discovered by the process when the sampling parameter is p_e. This means that at least one edge adjacent to the node fired. This probability is the complement of the node not being discovered, which means that none of the edges adjacent to the node fired when edges fire with probability p_e. Thus, conditioning on the node degree and sampling and sampling parameter of the adjacent edges, we have:

$$p_v = \sum_{k=0}^{\infty} f_D(k) \left[1 - (1-p_e)^k\right] \tag{3}$$

Similarly to $E[e(p_e)]$, we can then write $E[n(p_e)]$ as follows:

$$E[n(p_e)] = n_v p_v \tag{4}$$

We can establish limiting results for both $E[e(p_e)]$ and $E[n(p_e)]$ as p_e goes to 1. In particular, we can show that:

$$e^1 = \lim_{p_e \to 1} E[e(p_e)] = n_v E[D]/2 \tag{5}$$

$$n^1 = \lim_{p_e \to 1} E[n(p_e)] = n_v (1 - f_D(0)) \tag{6}$$

Thus, $p_e = 1$ reveals all the edges of G, but only its non-isolated nodes (i.e., nodes with positive degree).

As stated earlier, our ultimate goal is to establish a relationship between $E[e(p_e)]$ and $E[n(p_e)]$. In particular, can the edge sampling model lead to densification on the number of edges and nodes discovered? We start by defining two terms: densification and α-densification. The former means that the number of edges and nodes discovered are related super-linearly, but not necessarily through a power law relationship. The latter means that densification follows a power law relationship of the form $E[e(p_e)] \sim E[n(p_e)]^\alpha$ for an approximately constant $\alpha > 1$. It is usually understood that this relationship should span several orders of magnitude over $E[n(p_e)]$ to be unambiguous; we do not make this requirement explicit in the definition.

As we soon show, the edge sampling model can lead to densification, and under some conditions to α-densification. This indicates that densification can arise based on *how* we observe the fixed underlying graph, without requiring any dynamic growth process that modifies its structure.

We investigate the relationship between $E[e(p_e)]$ and $E[n(p_e)]$ by defining $\alpha(p_e)$, as follows:

$$\begin{aligned} \alpha(p_e) &= \frac{\partial \log\left(E[e(p_e)]\right)}{\partial \log\left(E[n(p_e)]\right)} \\ &= \left(\frac{\partial \log\left(E[e(p_e)]\right)}{\partial p_e}\right)\left(\frac{\partial \log\left(E[n(p_e)]\right)}{\partial p_e}\right)^{-1} \end{aligned} \tag{7}$$

Thus, $\alpha(p_e)$ denotes the instantaneous slope of the $E[n(p_e)]$ versus $E[e(p_e)]$ plot in log-log scale. Densification then means that $\alpha(p_e) > 1$ for several order of magnitude on $E[n(p_e)]$, while α-densification means that $\alpha(p_e)$ is approximately constant for several order of magnitude on $E[n(p_e)]$.

Using equations (2) and (4), we can derive $\alpha(p_e)$ analytically by applying its definition, which is given in equation (7). In particular, we have:

$$\alpha(p_e) = \frac{(1-p_e)\left(1 - \sum_k f_D(k)(1-p_e)^k\right)}{p_e \sum_k f_D(k) k (1-p_e)^k} \tag{8}$$

We can obtain both a lower bound and an upper bound for $\alpha(p_e)$. In particular we show that for $p_e > 1/2$, the following holds:

$$1 \le \alpha(p_e) \le \frac{1}{f_D(1)}. \tag{9}$$

The proofs are found in the appendix.

We can also establish two limiting results for $\alpha(p_e)$. Let α^1 be the limit of $\alpha(p_e)$ when p_e goes to 1. Similarly, let α^0 be the limit of $\alpha(p_e)$ when p_e goes to zero. We show that:

$$\alpha^1 = \lim_{p_e \to 1} \alpha(p_e) = \frac{1 - f_D(0)}{f_D(1)} \tag{10}$$

$$\alpha^0 = \lim_{p_e \to 0} \alpha(p_e) = 1 \tag{11}$$

Once again, the proofs are found in the appendix.

This result can be intuitively understood as follows. As p_e grows to 1, where the probability that an edge fires is quite large, the probability that a node of degree two or higher has been revealed is very close to one. The asymptotic slope of $\log E[n(p_e)]/\log E[e(p_e)]$ when almost all nodes have been discovered is then dominated by the discovery of the remaining edges and nodes of degree one, which occur at the same rate. As p_e shrinks to 0, where very few edges fire and therefore very few nodes are discovered, it is very likely that when an edge fires it will reveal two undiscovered nodes. Thus, the number of edges and nodes will grow linearly.

To support this last claim, we can establish yet another result for the case p_e is small, near zero. In particular, when p_e is small we have that $1 - (1-p_e)^d \simeq d p_e$. Therefore, $p_u \simeq p_e E[D]$. And finally,

$$E[n(p_e)] \simeq n_v E[D] p_e \tag{12}$$

Using equations (2) and (12) and applying it to equation (7), we obtain $\alpha(p_e) = 1$. Therefore, when p_e is sufficiently small, the discovery process through sampling does not yield densification.

It is interesting to note that the densification exponent, α, in the asymptotic regimes analyzed above (i.e., when p_e is near 0 or near 1), does not depend on the degree distribution, which may seem counter-intuitive. However, these results say nothing about the behavior of the $E[n(p_e)]/E[e(p_e)]$ curve for non-extreme values of the sampling parameter. In what follows, we estimate the densification exponent α for a range of values p_e when the degree of the underlying graph follows a specific distribution.

3.1 The power-law case

Motivated by the fact that many real network topologies exhibit a heavy tail node degree distribution [3, 1], we will consider a Zipf distribution to model the degree of the fixed underlying graph G. Thus,

$$f_D(k) = \frac{1}{k^s}\frac{1}{A}, \tag{13}$$

where s is a Zipf parameter and $A = \sum_{i=1}^{n_v} 1/i^s$ is the normalization factor with n_v corresponding to the total number of nodes in the underlying graph G. Note that as n_v grows large, the Zipf distribution approximates a power-law. Moreover, it can be shown that for very large n_v the distribution has infinite mean and variance for $s < 2$, finite mean and infinite variance for $2 \le s < 3$, and finite mean and variance for $s \ge 3$. From now on, we denote $c = \frac{1}{A}$.

We have:

$$1 = \sum_{k=1}^{n_v} c k^{-s} \sim \int_{k=1}^{n_v} c k^{-s}\, dk \quad \Longrightarrow \quad c = \frac{s-1}{1 - n_v^{1-s}} \simeq s - 1,$$

when $s > 1$ and n_v is large.

Given the sampling parameter p_e, the probability that a node with degree d is discovered is given by $1-(1-p_e)^d$. Note that this probability is close to 1 for large enough d. Moreover, when the node degree follows a Zipf distribution, nodes with large enough degree can exist with non-negligible probability. This motivates the following approximation for the probability that a node is discovered. If a node has degree greater than $1/p_e$, then with probability one it is discovered, otherwise with probability zero it is discovered. Thus, we let $1/p_e$ define a large enough degree. As before, let p_u denote the probability that a node is discovered. Using the approximation above, we have:

$$p_u \simeq P(D > 1/p_e) \tag{14}$$

where D is a random variable with distribution $f_D(k)$ that denotes the node degree.

Notice that out of the d edges incident to a given node with degree d, the expected number of edges that will fire is given by dp_e. The above approximation is equivalent to saying that a node will be discovered if this expectation if greater than 1, while it will remain unknown to the discovery process otherwise.

As before, assume the underlying graph has n_e edges and n_v nodes. Consider a sequence of values for p_e, in particular, let $p_e(i) = b^i, b < 1$ denote this sequence, for $i = 0, 1, \ldots$. Notice that when $i = 0$ all edges of the underlying graph are discovered (since $p_e(0) = 1$), while as we increase i, less and less edges are discovered. Let $e(i)$ and $n(i)$ denote the number of edges and nodes discovered at step i. As we showed:

$$E[e(p_e(i))] = n_e \cdot p_e(i)$$

$$E[n(p_e(i))] = n_v \cdot p_u(i),$$

where $p_e(i)$ and $p_u(i)$ correspond to the probability of edge and node discovery at step i.

For the number of discovered nodes, using the approximation above, we can write:

$$\begin{aligned} p_u(i) \simeq P(D > 1/p_e(i)) &\simeq \int_{k=b^{-i}}^{n_v} ck^{-s}\,dk \\ &\simeq \frac{c}{s-1} b^{i(s-1)} = b^{i(s-1)}. \end{aligned} \tag{15}$$

Thus, for n_v large, we obtain:

$$\begin{aligned} E[e(i)] &= n_e \cdot b^i \\ E[n(i)] &= n_v \cdot b^{i(s-1)}. \end{aligned}$$

Since n_e and n_v are fixed, in order to have $E[e(i)] \sim E[n(i)]^\alpha$ for all i, it suffices to have:

$$\alpha = \frac{1}{s-1}, \quad 1 < s < 2$$

This result establishes a direct relationship between the exponent of the degree distribution (s) of the underlying graph and the α-densification exponent. As we will soon show through numerical evaluations, this relationship is indeed a good approximation for the α-densification exponent.

Notice, however, that this approximation is too crude for $s > 2$, since $p_e(i)$ becomes larger than $p_u(i)$ in this case. For $s > 2$, we conjecture that α is close to 1, in particular for increasingly larger values of s. We will investigate this issue numerically in the next section. Thus, when $s > 2$ the edge sampling model does not yield densification.

3.2 Numerical Evaluations

In this section we conduct numerical evaluations of the edge sampling model presented above. We assume that the node degree of the underlying graph follows a Zipf distribution, given by equation ((13)), with s denoting the exponent. Moreover, we assume that the underlying graph has $n_v = 10^7$ nodes.

Figure 2 shows the evolution of $E[e(p_e)]$ and $E[n(p_e)]$ over different values for the sampling parameter p_e, computed using equations (2) and (4)), for a Zipf distribution with $s = 1.5$. Note that the y-axis in the bottom plot is in log-scale. The curves exhibit an interesting behavior. As shown in the top plot of Figure 2), at first, while p_e is small, the number of nodes discovered grows faster than the number of edges discovered. This occurs because in this range, almost every edge discovery results in the discovery of two nodes, which leads to a constant slope in edge-node curve. This is in line with our claim before, i.e. no densification when p_e is close to zero. However, as p_e increases, the number of nodes discovered grows much slower than the number of edges. This occurs because in this range almost all nodes have been discovered while edges continue to be discovered. Finally, as p_e approaches 1, the discovery process saturates and no new edges or nodes are discovered (bottom graph in Figure 2).

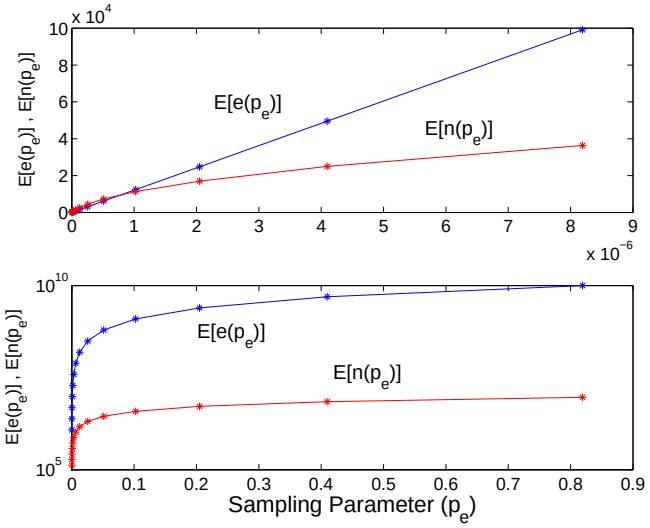

Figure 2: Evolution of the number of nodes ($E[n(p_e)]$) and the number of edges ($E[e(p_e)]$) discovered over change of sampling parameter for an underlying graph with a Zipf degree distribution ($n_v = 10^7$, $s = 1.5$, top plot is a zoom).

Figures 3 and 4 depict the expected number of nodes discovered versus the expected number of edges discovered using the same sampling parameter for Zipf distributions with different degree exponents. Notice that the plot is in log-log scale, thus, densification is present when the derivative (i.e., slope) of the curve is greater than 1 for a wide range of scales on $E[n(p_e)]$. Moreover, if this derivative is approximately constant for a wide range of scales on $E[n(p_e)]$, then the relationship between $E[e(p_e)]$ and $E[n(p_e)]$ is a power law of the type $E[e(p_e)] \sim E[n(p_e)]^\alpha$, where α is the slope of this straight line. Indeed, the curves for the Zipf distribution with $s < 2$ in Figure 3 show α-densification, as indicated by the theoretical analysis.

Figure 3 also illustrates the two different regimes for α. In particular, when p_e is small enough, we have α close to 1. As p_e

increases beyond this threshold, we have α greater than 1 and approximately constant. Moreover, as we increase s, the range over which densification is present (i.e., α greater than 1) decreases (for a fixed n_v), however, a higher α-densification exponent is achieved. For $s = 1.2$, α-densification spans less than two decades, while for $s = 1.5$ it spans three decades. Finally, for $s = 2.5$, we observe a slope of 1 for the entire range of values for p_e, i.e. no densification is present, as expected.

For comparison, we estimate the densification exponent after the threshold using linear regression of the data points in this interesting regime, denoted by $\hat{\alpha}$. We observe that the estimated slope found through numerical evaluations is close to our previous result, i.e. $\frac{1}{s-1}$. For example, for $s = 1.5$, we have that $\hat{\alpha} = 2$, which is equal to our theoretical approximation of $1/(s-1)$.

Figure 4 is simply a zoom for larger values of p_e, which corresponds to values that are close to the saturation of the discovery process (i.e., all nodes being discovered). In this regime, we expect the slope to be close to the limiting theoretical slope, α^1. For comparison, this slope is also shown in the plot for different values of s.

It is interesting to observe that α^1 –which is the asymptotic slope– is not far from $\hat{\alpha}$. Indeed, this indicates that graphs where nodes follow a Zipf distribution tend to saturate slowly, since they have many low-degree nodes. This results in an α-densification exponent that is close to $1/f_D(1)$.

We should again emphasize that the interesting regime is when the edges and nodes are being discovered, while we are away from both early transient and late saturation phases. In this range, we observe a constant densification exponent over wide range of $E[n(p_e)]$. For small sampling parameter, no densification is observed, while for large p_e we reach the asymptotic slope of α^1.

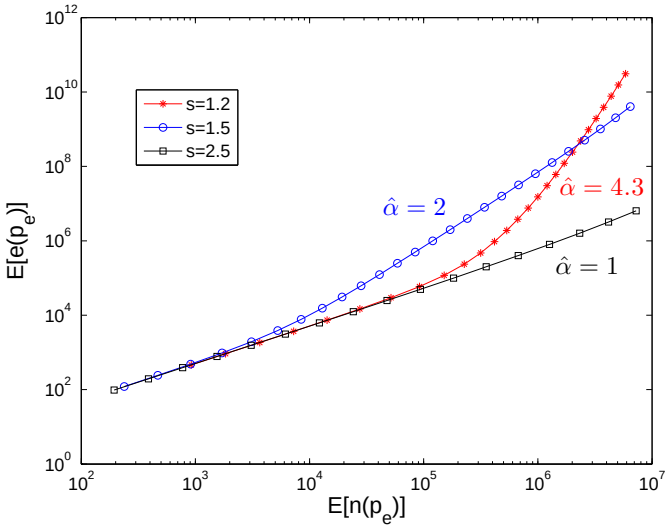

Figure 3: The number of nodes discovered ($E[n(p_e)]$) versus the number of edges discovered ($E[e(p_e)]$) by the same sampling parameter for Zipf degree distributions ($n_v = 10^7$, $s = 1.5$, $s = 1.2$, $s = 2.5$).

Figure 5 depicts the evolution of $\alpha(p_e)$ for Zipf distribution with different values of s, computed using equations (8). We first observe that densification occurs for all sampling parameters ($\alpha(p_e) > 1$ for all $0 < p_e < 1$). More surprisingly, we observe that $\alpha(p_e)$ quickly converges to a nearly constant value as p_e increases, and finally reach α^1. If this convergence occurs orders of magnitude

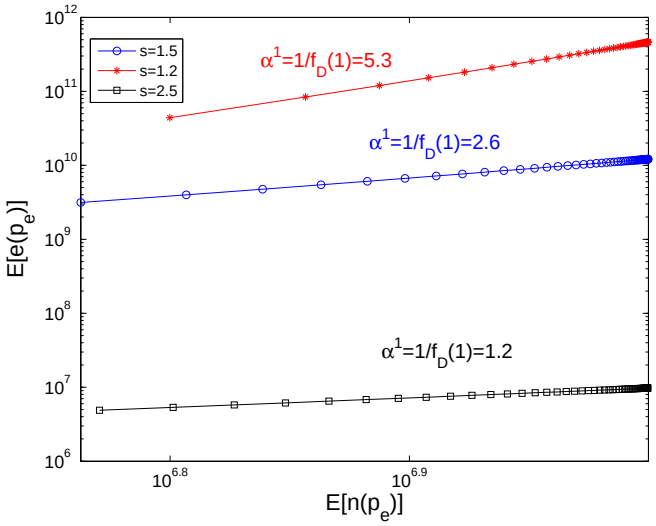

Figure 4: The number of nodes discovered ($E[n(p_e)]$) versus the number of edges discovered ($E[e(p_e)]$) by the same sampling parameter for Zipf degree distributions ($n_v = 10^7$, $s = 1.5$, $s = 1.2$, $s = 2.5$, zoom near larger $E[n(p_e)]$).

before saturation of the discovery process, then the model yields an α-densification. Finally, although α^1 depends only on $f_D(1)$ (see equation (10)), the underlying degree distribution plays a role on determining if the edge sampling model will yield α-densification.

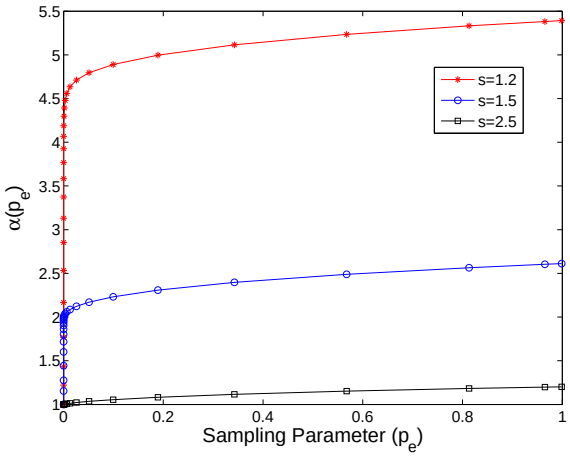

Figure 5: The densification exponent $\alpha(p_e)$ as a function of time for Zipf degree distribution ($n_v = 10^7$, $s = 1.5$).

Finally, in order to support our conjecture, Figure 6 depicts the relationship between the estimated densification exponent $\hat{\alpha}$ and the degree exponent s. As before, $\hat{\alpha}$ was computed through a linear regression of the points in the interesting regime (i.e., slope greater than 1). As illustrated in the figure, as s increases large, specifically when s is greater than 2, $\hat{\alpha}$ approaches 1. This supports our conjecture that α-densification occurs when $1 < s < 2$.

To summarize, the edge sampling model has a single parameter p_e that varies between zero and one and is used to sweep the edge-node curve. As p_e approaches zero, the slope of the curve approaches 1 and thus, no densification is present. However, for

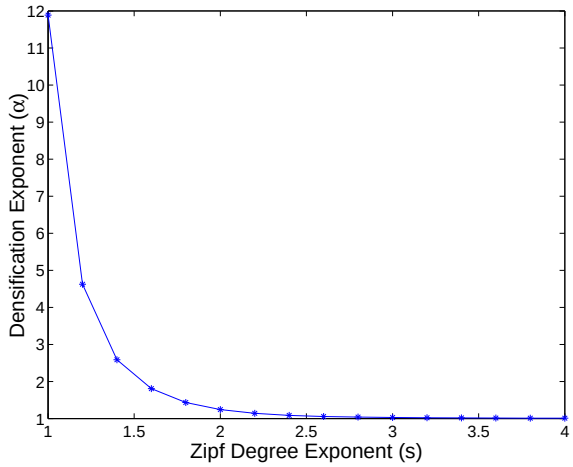

Figure 6: Relationship between the densification exponent (α) and the degree exponent (s) for an underlying graph with a Zipf degree distribution ($n_v = 10^7$)

sufficiently large p_e, it is possible that the curve follows a power-law, thus exhibiting α-densification. Moreover, this cut-off point for a sufficiently large p_e depends on the skewness of the degree distribution of the underlying graph. In particular, a larger skew (i.e., smaller s) requires a larger cut-off (i.e., a larger p_e). However, a larger skew also means that the densification exponent will be larger (since $\alpha = 1/(s-1)$). Finally, if the degree distribution is not skewed enough (i.e., if $s > 2$), then no cut-off exists, meaning that the slope is always 1, and thus, no densification is present.

3.3 Observation Models

The previous subsection presented the edge sampling model, which has a single parameter, namely, the edge sampling probability, p_e. In this section, we comment on the relationship between how datasets of real networks is gathered and the parameter p_e. In particular, considering how these datasets are obtained, we believe that the parameter p_e varies for either of the following reasons:

Accumulation In this interpretation of the sampling model, knowledge about the edges and nodes of the underlying graph is simply accumulated over time. This captures studies where each new edge (and consequently nodes) are added to an evolving graph; this is the case, for example, in arXiv citation graph, or IMDB actors-to-movies graph (analyzed in [9]). Thus, the key parameter of this model is time, which will determine the probability that edges and nodes are discovered, i.e. $p_e = f(t)$. As time grows from 0 to ∞, the sampling parameter varies from zero to one, resulting in the gradual discovery of edges and nodes.

We suppose any edge e is associated with an independent renewal process of rate λ, started in steady state. We define then $F_F(t, \lambda)$ to be the inter-sampling time cumulative distribution for an edge with a sampling rate of λ. Moreover, the sampling rate of an edge is also a random variable with cumulative distribution $F_R(\lambda)$, identical for all edges of G. We will also assume that sampling rates are chosen independently and that edges are also sampled independently of each other.

Modulation In this interpretation of the model, a *fixed time window* is considered. In other words, the sampling parameter, p_e, is a function of a global sampling rate, which determines the rate with which the edges fire within a given fixed length time window. Different time windows can have different intensities with respect to the edge sampling rates. This is motivated by the fact that analysis of real data is often done considering fixed windows of time over the dataset, as opposed to an increasing time interval. However, different windows of time may have different sampling rates. For example, consider the study of email networks and a time window of one month. It is natural that different months will have different intensities of email exchange. For example, a vacation month is surely to have a lower sampling rate than a end-of-semester month, when considering the email network of a large university.

As before, we introduce an inter-sampling time distribution. However, here we define a modulating intensity parameter Λ, which influences the edge sampling rates. So we have $p_e = f(\lambda)$. In particular, the edge sampling rate will be given by:

$$\lambda = \Lambda\lambda_o \tag{16}$$

where λ_o is a random variable that follows the original edge sampling rate cumulative distribution, $F_R(\lambda)$. Note that the actual edge sampling rate distribution is just the original edge sampling rate distribution scaled by a constant Λ.

Notice that we are only giving different interpretations to p_e. In particular, we let p_e be controlled either by t or by Λ, which will be the case when we apply the edge sampling model to real datasets.

4. EDGE SAMPLING AND REAL DATA

In this section, we apply the edge sampling model to two datasets of real network data. The goal is to illustrate that the sampling model can capture and thus partially explain, the observed densification when considering data from these networks. In order to do this, we consider the two variations of the model, i.e. accumulation and modulation. Thus, the sampling parameter p_e will be a function of time or intensity, respectively, and is determined through the choice of this function, as described below.

Unfortunately, we do not have the real underlying graph to drive the edge sampling process. We also have no knowledge of the edge sampling probability, p_e, which can be characterized by the edge sampling rates and the inter-sampling time distribution. Therefore, in order to apply the model we propose, we will use the actual dataset to derive estimates for these unknown parameters.

The underlying graph $G = (V, E)$ we consider is given by the accumulated graph over the entire dataset, that is

$$G[0, t_f] = (V[0, t_f], E[0, t_f]),$$

where t_f is the length of the available dataset. Thus, we will assume that the degree distribution of the underlying graph is given by the empirical degree distribution of $G[0, t_f]$, namely $f_D(k)$.

Concerning the edge sampling process, we will assume that the edge sampling rate is given by its average over the entire dataset. Thus, for each edge $e \in E$, we define $\lambda_e = |M_{ij}[0, t_f]|/t_f$, where edge $e = (i, j)$ and $M_{ij}[0, t_f]$ is the set of samples of the edge (i, j) available in the dataset in the period $[0, t_f]$. Using the sampling rates λ_e for all $e \in E$, we can determine the empirical edge sampling rate distribution, namely, $f_R(\lambda)$. Finally, we assume that the inter-sampling times of an edge $e \in E$ is exponentially distributed with an associated rate λ, i.e. the sampling parameter p_e will be a function of time. Thus, $F_F(t, \lambda) = p_e(t) = 1 - e^{-\lambda t}$.

We will consider EPFL email dataset described in Section 2. Recall that this dataset leads to α-densification, as illustrated in Figure 1. Using the same dataset, we obtain the necessary parameters to construct the edge sampling model, as described above. Finally, we numerically compute $E[n(t)]$ and $E[e(t)]$ for this specific model. As with the original dataset, t is measured in weeks and varies from 1 to 89.

The results produced by the model are shown in Figure 7. The plot clearly indicates that the edge sampling process leads to densification on the number of nodes and edges that are discovered over time. Moreover, the densification produced by the model seems to follow a power law, with $\hat{\alpha} = 1.8$ (recall that $\hat{\alpha}$ is the slope of the line obtained through a linear regression over the points in the plot). Surprisingly, this exponent closely matches the actual data ($\hat{\alpha} = 1.57$, see Figure 1) which indicates that the edge sampling process is a very plausible explanation for the observed densification.

To assess the goodness of the fits, we consider the confidence interval for the slope of the line obtained through linear regression. Using simple statistical tools, we obtain the 95% confidence interval for the slope of the fitted line. For EPFL email data set, we find a confidence interval of $(1.55, 1.59)$ with the slope being $\hat{\alpha} = 1.57$. Thus, the estimated slope is well above 1 with high probability, which shows a clear densification as the network grows. The same is true when applying the edge sampling model to real data, giving a slope of sufficiently larger than 1 with small error.

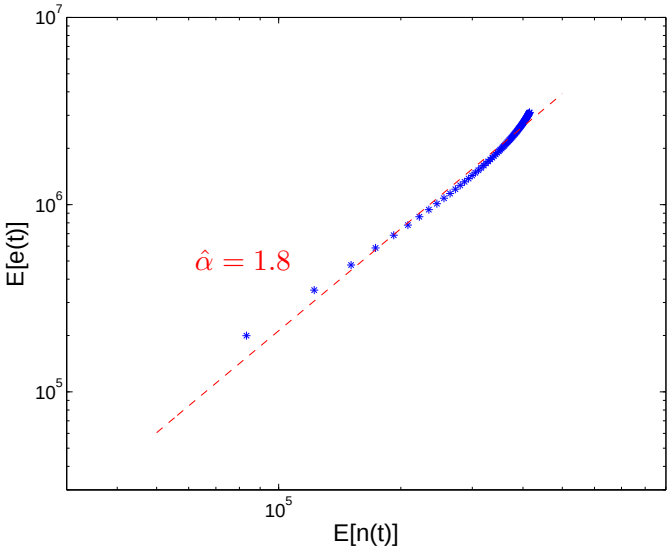

Figure 7: The edge sampling model applied to the EPFL email network dataset.

4.1 Modulation Model

In this section we will apply the edge sampling model over a fixed window of time (proposed in Section 3.3 as *modulation*) to real data. Recall that the motivation for this model is that real data is often provided in snapshots over a fixed window of time. Moreover, even when this is not the case, data analysis often considers data only over a fixed window of time, varying the position of the window over the dataset.

4.1.1 EPFL Email Dataset

We start by presenting the results for the actual EPFL email dataset. Recall that EPFL email dataset is provided in snapshots of weeks (i.e., all messages that were sent/received during a given week). Thus, for each week $t = 1, \ldots, 89$, we can define the graph $G[t, t] = (V[t, t], E[t, t])$, which is defined using only messages exchanged during that week. For each graph, we let $n(t) = |V[t, t]|$ denote the number of nodes and $e(t) = |E[t, t]|$ denote the number of edges. Differently from before, notice that $n(t)$ and $e(t)$ correspond to the number of nodes and edges respectively, seen on week t only.

Figure 8 shows the plot of $n(t)$ versus $e(t)$ for $t = 1, \ldots, 89$ for EPFL email dataset. Interestingly, we again observe densification despite the fact that information about the graph is not accumulated over time. Thus, densification arises here for reasons other than accumulation over time, since each point corresponds to exactly one week. Moreover, there is no trend between time and the points in the plot. Although the points do not form a straight line, there is a clear increasing trend among them. When fitted to a straight line, we obtain a slope of $\hat{\alpha} = 1.89$, as illustrated in the figure.

As before, we find the confidence interval for the estimated slope, which in this case is $(1.65, 2.12)$. Although not being tight bounds, we can still claim that with high probability we observe a clear densification.

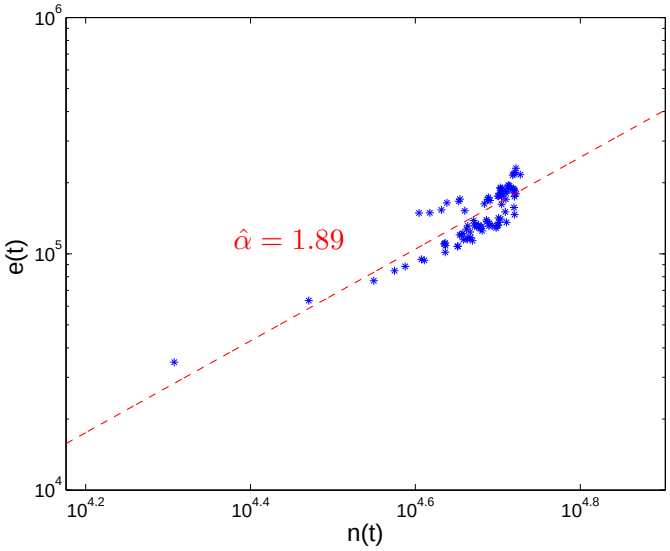

Figure 8: Densification of EPFL email dataset using a fixed time window (1 week).

In order to apply the edge sampling model through modulation we again need to determine its parameters. Besides the underlying graph and the sampling process, we also need to determine the edge sampling intensity. We will estimate the actual underlying graph and sampling process as before, by considering the graph accumulated over the entire dataset and the average rate of messages per edge. However, we still need to determine the edge sampling intensity for each fixed time window.

Let Λ_t denote the edge sampling intensity of week t. We define $\Lambda_t = |M[t, t]|/\Lambda$, where $M[t, t]$ is the set of messages in the dataset during week t and Λ is the overall average message rate (per week). In particular, we define $\Lambda = |M[1, 89]|/89$. Thus, Λ_t measures the average intensity of the edge sampling rate of week t in relationship to the overall average edge sampling rate. In other words, the sampling parameter p_e will be a function of the modulating intensity parameter in this case.

We can now apply the edge sampling model with fixed time

window. Notice that each week t corresponds to a different time window with edge sampling intensity Λ_t. Figure 9 shows the expected number of nodes $E[n(\Lambda_t)]$ versus the expected number of edges $E[e(\Lambda_t)]$ discovered for each week t. It is clear that the model yields densification when considering the different weeks. Moreover, there is strong linear dependence between $E[n(\Lambda_t)]$ and $E[e(\Lambda_t)]$ indicating a power law relationship. Finally, a straight line fitted to the data yields the slope $\hat{\alpha} = 1.5$, which is fairly close to the actual dataset, with a confidence interval far above 1. This supports the claim that the proposed edge sampling model captures the discovery process of edges and nodes.

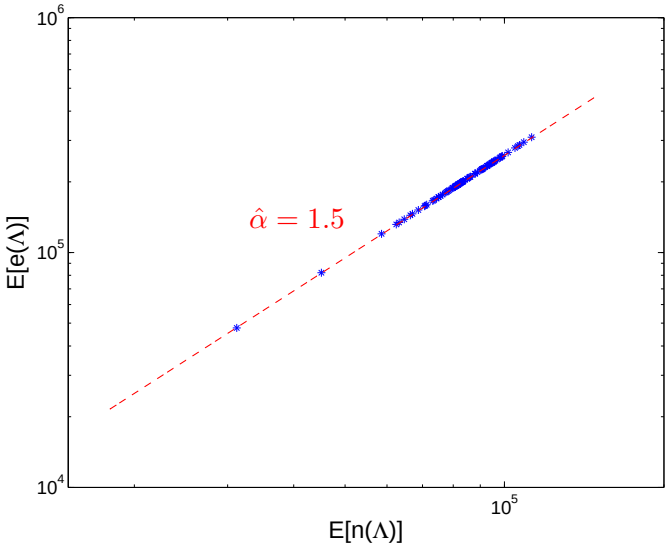

Figure 9: The edge sampling model with fixed time window applied to EPFL email dataset (1 week).

4.1.2 AS Graph Dataset

In this section we consider another dataset of network data that is publicly available, the AS graph. The Internet today is composed of several thousands Autonomous Systems (ASes) that interconnect with each other providing global connectivity. Most of ASes are owned and operated by Internet Service Providers (ISPs) like AT&T or MCI, while other ASes belong to smaller businesses or universities. The interconnection of ASes forms a graph, which is known as the AS graph.

Information about the AS graph is provided daily through snapshots generated from an aggregate of information about the network. However, for a myriad of reasons, these daily snapshots provide only an estimate of the real AS graph. Moreover, the real AS graph is constantly changing as both ASes and their interconnections are added and deleted over time.

We consider a dataset of the AS graph formed by 735 daily snapshots. As for EPFL email dataset, each snapshot defines a graph $G[t,t] = (V[t,t], E[t,t])$ formed by all nodes and edges that appear in day t. For each day t, we thus have a number of nodes $n(t) = |V[t,t]|$ and a number of edges $e(t) = |E[t,t]|$.

Figure 10 shows the result of plotting $n(t)$ versus $e(t)$ in log-log scale for all days. Although the number of nodes and edges vary little from one day to the other, the plot still strongly indicates a power law densification with $\hat{\alpha} = 1.09$.

To assess the goodness of this fit, we calculate the confidence interval for the slope, which yields the interval $(1.08, 1.1)$. Thus although $\hat{\alpha}$ is close to 1, with high probability it will be a value above 1, resulting in densification.

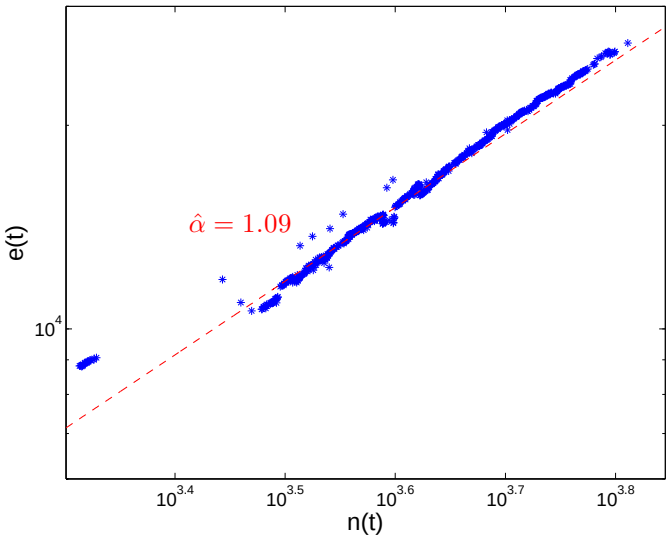

Figure 10: Densification of the AS graph dataset using a fixed time window (1 day).

We will apply the edge sampling model with fixed time window to the AS graph dataset. Notice that since each snapshot provides full view of the AS graph, the edge sampling model where information is accumulated over time is not suitable for this dataset. However, in order to apply the fixed time window model we need to estimate its parameters from this dataset. We execute the same procedure used for EPFL email dataset, estimating the fixed underlying graph and the edge sampling rates using the accumulated graph[2]. We also determine, for each daily snapshot, the edge sampling intensity in the same manner.

Figure 11 shows the result obtained with the model for the various sampling intensities Λ_t. Once again, we observe densification on the number of nodes and edges discovered when considering each fixed time window. Moreover, there is a roughly linear trend in this relationship, indicating a power law densification, with $\hat{\alpha} = 2$, as illustrated in the plot.

Although the slope of the fitted line ($\hat{\alpha}$) yielded by the model is not very close to that of the actual data, we still believe the model is representative of how nodes and edges are revealed in this dataset, and a clear densifying behavior is observed.

There can be different reasons why the model does not match the data accurately. First, the real underlying graph is unknown, and it also changes in time. We made the assumption of having the degree distribution of the final accumulative graph for the underlying graph in order to apply the model to real data and do the simulations. As seen before, the degree distribution of the underlying graph is a key factor in the densification behavior and the observed exponent. Thus, this assumption surely affects the result when edge sampling is applied to real data. Moreover, in applying the model to fixed time windows, we estimated an average intensity through finding the ratio of total number of messages in a fixed window and the total number of messages in the final accumulative graph. How accurate this estimation is still needs to be investigated on more

[2]In the AS graph dataset, for each snapshot, each edge has either rate 1 (i.e., edge is present in the AS graph) or 0 (i.e., edge is not present).

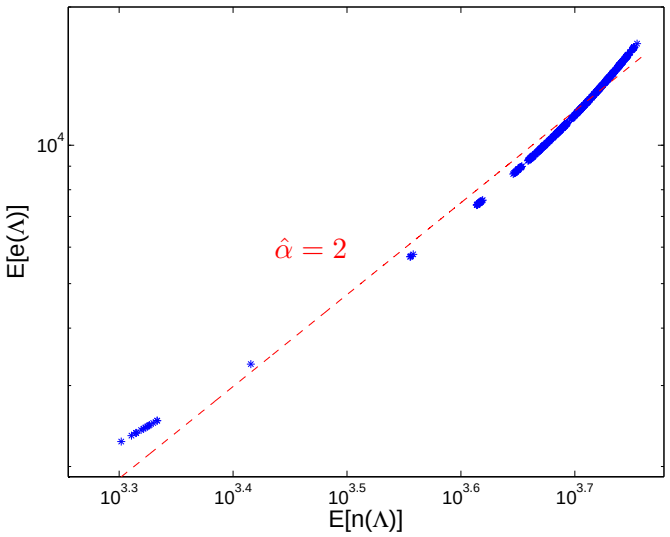

Figure 11: The edge sampling model with fixed time window applied to the AS graph dataset 1 day).

datasets. Summing up, although not matching the data accurately for the reasons above, but we observe a neat densification behavior applying the edge sampling model to the data, finding exponents not far from reality.

We have also applied the proposed models to the exact structure of the underlying graph (where the underlying graph is given by the final cumulative graph), as opposed to considering only the empirical degree distribution. Using the edge sampling model with exponential inter-sampling time distribution, we observed that results from the models match real data very accurately for both cases of accumulation and modulation. This further strengthens our claim that densification arises regardless of the process used for sampling the edges.

We also considered a couple other datasets analyzed by Leskovec et al. [9], namely the IMDB actors-to-movies bipartite graph and arXiv citation graph. However, these datasets are not good examples for which the proposed edge sampling model can be applied. To be more precise, in the citation graph, nodes (i.e., papers) in the graph appear first, and edges (i.e., citations) appear afterwards, when a given paper cites another. Thus, in this dataset, new nodes are not revealed through sampling the edges of the graph. In the IMDB dataset, movies and actors correspond to the two different types of nodes of a bipartite graph. An edge between a given pair of nodes is present when a given actor plays a role in a given movie. Thus, each edge is sampled only once (i.e., when an actor plays in a new movie), as the edge will never be sampled again, giving rise to a constant edge sampling rate for all edges. For these reasons, in this work, we do not consider these datasets in the framework of the proposed edge sampling model.

5. RELATED WORK

Network (or graph) densification is a phenomenon that has been recently observed by Leskovec et al. in various datasets of real data over time [8, 9]. In particular, they empirically observe that the number of edges was growing much faster than the number of nodes in these graphs. More surprising, for all datasets considered, this relationship seemed to follow a power law of the type

$$e(t) \propto n(t)^a,$$

where $e(t)$ and $n(t)$ denote the number of edges and nodes of the graph at time t, and a is called the densification exponent, which is a constant greater than 1. Under this relationship, known as *densification power law*, the average node degree grows with time and, thus, the graph densifies.

Most of the graphs considered in their work were generated by accumulating the data available in the datasets over time. However, they also considered the case where the graph is generated by considering only a fixed window of time. This is the case for the AS graph and the email network. For the email network, they observe that the densification exponent is larger for the fixed time window case, which is consistent with our observations.

Besides providing empirical evidence of network densification, Leskovec et al. also provide mathematical models that can capture and partially explain this phenomenon [8, 9]. The intuition behind their models is that densification occurs due to the growth of the network. Basically, nodes created at a later point tend to establish mode edges than nodes created earlier. The proposed *Forest Fire Model* is based on this intuition and is a model for network growth, where nodes and edges are added to the graph over time. They also prove that this model leads to densification of the network over time. Finally, Leskovec et al. also investigate other growth models for networks, showing how properties such as power law densification can arise, in particular when using the model of Kronecker Graphs [6, 7].

An observation similar to the findings of Leskovec et al. concerning network densification was also previously made by Dorogovtsev and Mendes [4]. They empirically observed that the graph formed by the World Wide Web was densifying over time, naming this phenomenon *accelerated growth*. They also propose a model to capture this phenomenon where the average node degree grows as a power law in time.

6. CONCLUSION

This work investigates the recently observed phenomenon known as *densification*, where the number of edges of a network grows much faster than the number of nodes. In particular, we provide a novel explanation for this phenomenon when considering some real datasets. The key idea is that densification arises naturally from the process used to reveal the edges and nodes of the unknown underlying graph. For most datasets, this process is based on the discovery of edges, that are then used to discover nodes. Based on this idea, we propose the *edge sampling model*, where edges from a fixed underlying graph are sampled by changing a sampling parameter p_e, leading to their discovery and to the discovery of adjacent nodes.

By assuming a heavy-tailed degree distribution for the underlying graph, we prove properties concerning the densification of the number of edges and nodes discovered over different values for the sampling parameter. In particular, we show that a power law relationship, which we call α-densification, can arise over a wide range of scales with exponent given by $\frac{1}{s-1}$, where s is the Zipf exponent of the node degree distribution of the underlying graph. We also prove limiting results for densification exponent as p_e approaches both extremes (i.e., 0 and 1), and show that densification can be present and it is bounded for all $0 < p_e < 1$.

We also comment on the relationship of empirical network studies to the edge sampling model, and specifically, on what factors affect the sampling probability p_e. We introduce two observation models, namely *accumulation* and *modulation*. In the first variation, edges and nodes are discovered and accumulated over time, while in the second variation time is fixed but the edge sampling intensity for a given time window is allowed to vary.

Finally, we consider datasets of real graphs, namely EPFL email

network and the Internet AS graph, both of which show densification over time. We apply the two variations of our model to these datasets and the results obtained indicate that the edge sampling model is indeed a plausible alternative explanation for the observed densification phenomenon.

Acknowledgments

We gratefully acknowledge the extensive help we have received from EPFL's postmaster, Martin Ouwehand. His assistance with harvesting and interpreting the email dataset was invaluable. We also extend our thanks to Leskovec, Kleinberg and Faloutsos, who made available some of the datasets analyzed in their seminal work [9].

7. REFERENCES

[1] R. Albert and A. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[2] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks : Structure and dynamics. *Physics Reports*, 424(4-5):175–308, feb 2006.

[4] S. N. Dorogovtsev and J. F. F. Mendes. Accelerated growth of networks. In *Handbook of Graphs and Networks: From the Genome to the Internet, S. Bornholdt and H.G. Schuster*, Eds, Wiley-VCH, Berlin, Germany, 2002.

[5] J. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: measurements, models and methods. In *Proc. of the 5th International Computing and combinatorics Conference (COCOON)*, 1999.

[6] J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos. Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In *European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, 2005.

[7] J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using kronecker multiplication. In *International Conference on Machine Learning (ICML)*, 2007.

[8] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2005.

[9] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (ACM TKDD)*, 1(1), 2007.

[10] M. Newman. The structure and function of complex networks. *SIAM Reviews*, 45(2):167–256, Mar. 2003.

APPENDIX

A. PROOFS

Here we find upper and lower bounds for equation (8), i.e.

$$\alpha(p_e) = \frac{(1-p_e)\left(1-\sum_k f_D(k)(1-p_e)^k\right)}{p_e \sum_k f_D(k)k(1-p_e)^k}$$

We first consider the upper bound. For $p_e > 1/2$, we have $1-p_e < p_e$. Thus we can write:

$$\frac{(1-p_e)\left(1-\sum_k f_D(k)(1-p_e)^k\right)}{p_e \sum_k f_D(k)k(1-p_e)^k} \overset{(a)}{\leq} \frac{1-p_e}{\sum_k f_D(k)k(1-p_e)^k} = \frac{1}{f_D(1)+f_D(2)(1-p_e)+\cdots} \leq \frac{1}{f_D(1)},$$

where (a) follows since the second term in the numerator is $P(v)$, and thus less than 1.

Now we consider the lower bound. We can rewrite $\alpha(t)$ as:

$$\alpha(t) = \underbrace{\left(\frac{1-p_e}{\sum_k f_D(k)k(1-p_e)^k}\right)}_{Q}\underbrace{\left(\frac{1-\sum_k f_D(k)(1-p_e)^k}{p_e}\right)}_{T}. \tag{17}$$

Assuming $f_D(0) = 0$ (which is the case for Zipf distribution), and using $\sum_k f_D(k) = 1$, we have,

$$Q = \frac{f_D(1)(1-p_e)+f_D(2)(1-p_e)+f_D(3)(1-p_e)+\cdots}{f_D(1)(1-p_e)+f_D(2)\cdot 2e(1-p_e)^2+\cdots} \tag{18}$$

Comparing the coefficients of $f_D(k)$ in the numerator and denominator of (18) for all $k > 1$, we have,

$$(1-p_e) \geq k(1-p_e)^k, k = 2,3,\ldots \quad \text{if } \ln\frac{1}{1-p_e} \geq \frac{\ln(k)}{k-1}. \tag{19}$$

Since $\frac{\ln(k)}{k-1}$ is a decreasing function of k, the condition $p_e \geq 1/2$ (derived by substituting $k = 2$) suffices to have each term in the numerator equal or greater than the corresponding term in the denominator, i.e.

$$f_D(k)(1-p_e) \geq f_D(k)k(1-p_e)^k, k = 1,2,3,\ldots \quad \text{if } p_e \geq 1/2, \tag{20}$$

which results in $Q \geq 1$. Note that for $k = 1$ the two corresponding terms are equal, thus satisfying (20).

Following the same approach for T, we have,

$$T = \frac{1-\overbrace{\left(f_D(1)(1-p_e)+f_D(2)(1-p_e)^2+\cdots\right)}^{Y}}{1-\underbrace{\left(f_D(1)(1-p_e)+f_D(2)(1-p_e)+\cdots\right)}_{Z}}. \tag{21}$$

Comparing the coefficients of $f_D(k), k = 1,2,\ldots$ in Y and Z, we observe,

$$f_D(k)(1-p_e)^k \leq f_D(k)(1-p_e), k = 1,2\ldots, \tag{22}$$

which results in $Z \leq Y$, yielding $T \geq 1$. Thus,

$$\alpha(t) = Q \cdot T \geq 1, \text{ if } p_e \geq 1/2, \tag{23}$$

which proves the lower bound in equation (9).

Putting all together, we proved that for $p_e \geq 1/2$,

$$1 \leq \alpha(p_e) \leq \frac{1}{f_D(1)}. \tag{24}$$

Next, we prove the two limiting results for $\alpha(p_e)$ as defined in equation (8). Define α^0 to be the limit of $\alpha(p_e)$ when p_e goes to

zero. Thus,

$$\begin{aligned}\alpha^0 &= \lim_{p_e\to 0}\alpha(p_e)\\ &= \lim_{p_e\to 0}\frac{(1-p_e)\left(1-\sum_k f_D(k)(1-p_e)^k\right)}{p_e\sum_k f_D(k)k(1-p_e)^k}\\ &= \lim_{p_e\to 0}\frac{1-\sum_k f_D(k)(1-pk+\cdots)}{p\sum_k f_D(k)k(1-pk\cdots)} = 1\end{aligned}$$

where we used the Taylor's expansion for $(1-p)^k$.

Now define α^1 to be the limit of $\alpha(p_e)$ when p_e goes to 1. Thus,

$$\begin{aligned}\alpha^1 &= \lim_{p_e\to 1}\alpha(p_e)\\ &= \lim_{p_e\to 1}\frac{(1-p_e)\left(1-\sum_k f_D(k)(1-p_e)^k\right)}{p_e\sum_k f_D(k)k(1-p_e)^k}\\ &= \lim_{p_e\to 1}\frac{1-\sum_k f_D(k)(1-p_e)^k}{p_e}\cdot\frac{(1-p_e)}{\sum_k f_D(k)k(1-p_e)^k}\\ &= \lim_{p_e\to 1}\frac{1-\sum_k f_D(k)(1-p_e)^k}{p_e}\cdot\frac{1}{\sum_k f_D(k)k(1-p_e)^{k-1}}\\ &\overset{(b)}{=} (1-f_D(0))\left(\frac{1}{f_D(1)}\right)\\ &= \frac{1-f_D(0)}{f_D(1)} \qquad (25)\end{aligned}$$

where (b) follows since in the first term of the product, the denominator goes to one and all terms in the sum go to zero except for $k=0$, and in the second term of the product all terms in the sum go to zero except for $k=1$.

Note that when $f_D(0)=0$, equation (25) simplifies to $\frac{1}{f_D(1)}$. This simplification applies in the case of the Zipf distribution, where $f_D(0)=0$ independent of its parameter s and n. Moreover, for the Zipf distribution, we have that $f_D(1)=1/A$, where A is the normalization factor of the distribution, which is given by:

$$A=\sum_{i=1}^{n}1/i^s = 1+\frac{1}{2^s}+\frac{1}{3^s}+\cdots+\frac{1}{n^s}. \qquad (26)$$

Thus, for the Zipf distribution, we have that $\alpha^1 = 1/f_D(1) = A$.