Modeling Network Growth under Resource Constraints*

ABSTRACT

We propose a resource-constrained network growth model that explains the emergence of key structural properties of real-world directed networks: heavy-tailed indegree distribution and high local clustering and the degree-clustering relationship. In real-world networks, individuals form edges under constraints of limited network access and partial information. However, well-known growth models that preserve multiple structural properties do not incorporate these resource constraints. Conversely, existing resource-constrained models do not jointly preserve multiple structural properties of real-world networks. We propose a random walk growth model that explains how real-world network properties can jointly arise from edge formation under resource constraints. In our model, each node that joins the network selects a seed node from which it initiates a random walk. At each step of the walk, the new node either jumps back to the seed node or chooses an outgoing or incoming edge to visit another node. It links to each visited node with some probability and stops after forming a few edges. Our experimental results against four well-known growth models indicate improvement in accurately preserving structural properties of five citation networks by a factor of 2 to 3. Our model also preserves 2 structural properties that most growth models cannot: skewed local clustering distribution and bivariate indegree-clustering relationship.

KEYWORDS

Network growth models, Network evolution, Network structure, Resource Constraints

ACM Reference Format:
. 2017. Modeling Network Growth under Resource Constraints. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

We develop a resource constrained model of network growth that explains the emergence of key structural properties. The problem is important for several reasons. Individuals form real-world networks acting under resource constraints and while using local information. These networks that individuals form exhibit rich structural properties. However, we lack an understanding of mechanisms that are resource constrained and which use local information explain the emergence of these structural related properties.

Classic models of network growth, make unrealistic assumptions about what agents who form edges do. Consider as a simple stylized example, the process of finding the a set of papers to cite when writing an article. In the preferential attachment model [3] of network growth, a node making m citations would pick a paper uniformly at random from *all* papers in the domain, and either cite it or copy one of its references. We would repeat this process, till we've exhausted our budget of m references. Notice that the process assumes access to the entire dataset, and that one would pick papers uniformly at random. An equivalent formulation of this "vertex copying" model is to cite papers from the entire dataset in proportion to their in degree. The latter formulation assumes that agent making citations know the entire in degree distribution. While preferential attachment model explains well the emergence of the power-law degree distribution, the attachment model is an unrealistic representation of how agents make decisions on edge formation.

The problem of developing a model of network growth, where agents act under resource constraints, including access to only local information is hard. The problem lies in identifying simple mechanisms, with few parameters, where the agents only use local information and *jointly* preserve the properties related structure.

We propose a random walk based model of network growth that jointly explains the emergence of the following properties: heavy-tailed in-degree distributions, local clustering and clustering-degree relationships. In the growth model, an incoming node picks a recent node as the seed. It will link to this node with some constant linking probability. Then, it follows the outgoing link or the incoming link of this seed node and arrives at a new node. At each new node, it decides to link with the same constant linking probability. Then it has to decide whether to jump back to the seed node, or following incoming or outgoing links. The process repeats until the agent has exhausted its budget for linking.

Our main contributions are as follows:

- We propose a resource constrained model of network growth using a local edge formation process.
- We propose a model that jointly explains multiple structural properties, including in-degree distribution, clustering, degree clustering relationship and edge densification.

We conducted extensive experimental results, against state of the art baselines, on large citation datasets. We show that our growth model outperforms that best competing model in preserving the joint structural properties—degree distribution, clustering and degree-clustering relationship—by a margin of 73.56% averaged over all datasets.

The rest of the paper is organized as follows. In Section 2, we define key structural properties and introduce the datasets. In Section 3, we formally state the goal of this paper. Then, in Section 4,

*Produces the permission block, and copyright information

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

we provide an overview of existing network models. This is followed by Section 5, where we report prominent structural characteristics of citation networks. In Section 6, we propose a resource-constrained growth model. We validate our model in Section 7. Finally, we present the related work in Section 8.

2 PRELIMINARIES

In this section, we first define important structural characteristics that describe network structure. Then, in Section 2.2, we describe the network datasets used in this paper.

2.1 Structural Properties

Now, we discuss four well-known structural properties: degree distribution, local clustering coefficient, the relationship between degree & local clustering and average path length. These properties are widely used [2], compact statistical descriptors of network structure.

The degree distribution of an undirected graph is the probability distribution $p(k)$ of nodes with degree k. With directed graphs, we can compute the degree distribution separately for indegree and outdegree. The well-known pagerank centrality measure has positive correlation with indegree [9]. Therefore, indegree distribution indicates how node centrality distributed in directed networks.

The local clustering coefficient of a node measures the edge density of the node's neighborhood. For example, the clustering coefficient of an individual in an undirected social network is the ratio of observed friendships amongst neighbors to all possible friendships amongst neighbors. In directed networks, the neighborhood of node v_i can refer to the set of nodes that link to v_i, set of nodes that v_i links to or the union of both. In this paper, we define the neighborhood of v_i to be a set of all nodes that link to v_i. More formally, the local clustering coefficient C_i of node v_i with neighborhood N_i and indegree k_i in a directed network $G = (V, E)$ is defined as follows.

$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}$$

This equation states that the local clustering coefficient of v_i in a directed network is the number of observed directed relationships divided by the maximum possible directed relationships in the neighborhood of v_i.

The bivariate relationship between degree and local clustering coefficient is important. This property sheds light on the variation of node neighborhood density as a function of node degree. In real directed networks, average local clustering decreases as indegree increases [26].

The average path length is the expected length of the shortest path between two randomly picked nodes in a network. First studied by Milgram [22] and subsequently validated by experiments [18], large real networks tend to have small average path length.

We use these properties to uncover common traits in the structure of real networks and empirically validate the effectiveness of our proposed model in Section 5 and Section 7 respectively.

2.2 Datasets

In this paper, we consider five citation networks. Citation networks are directed networks in which nodes are papers and edges are citations from one paper to another.

We focus on citation networks for three reasons. First, nodes form all edges to existing network nodes at the time of joining the citation network. Since nodes do not form or delete edges at a later time, citation networks allow us to carefully analyze how new nodes that join the network form edges. Edge dynamics such as the deletion and addition of edges are important and we plan to investigate them at a later time. Second, citation network datasets include the time (e.g. publication year) at which papers join the network. Therefore, structural properties can be better understood by studying network "snapshots" at different stages of the growth process. Third, citation networks are large networks that span many years. As a result, the structural properties, defined in Section 2.1, are distinct and well-defined.

We consider the citation networks of academic papers, patents and judicial cases; Table 1 provides the basic statistics of these networks:

- **ArXiv HEP-PH** (HEP-PH) [11] is an academic citation network of HEP-PH (high energy physics phenomenology) papers in the ArXiv e-print.
- **U.S Patents** (PATENTS) [19] is a citation network of U.S. utility patents maintained by the National Bureau of Economic Research.
- **APS Journals** (APS) [1] is an academic citation network maintained by the American Physical Society (APS) that consists of articles published in APS journals.
- **Semantic Scholar** (SEMANTIC) [2] is a citation network of all Computer Science and Neuroscience papers made public in June 2017 by Semantic Scholar, an academic search engine corpus.
- **U.S. Supreme Court Cases** (USSC) [10] is a citation network in which nodes are U.S. Supreme Court cases. There is an edges from case i to case j if and only if case i cites case j in its majority opinion.

Table 1: Dataset statistics. We report the statistics of five real-world datasets from various domains—United States Supreme Court Cases, patents and academic publications such APS and HEP-PH—used in our experiments.

Network	Nodes	Edges	Time range
USSC	30,228	216,738	1754-2002
HEP-PH	34,546	421,533	1992-2002
APS	577,046	6,967,873	1941-2015
Patents	3,923,922	16,522,438	1975-1999
Semantic	7,706,506	59,079,055	1991-2016

In this section, we reviewed key structural properties that network growth models try to preserve. Then, we briefly described

[1] https://journals.aps.org/datasets

[2] http://labs.semanticscholar.org/corpus/

the citation network datasets that we use in this paper. In the next section, we describe the main edge formation mechanisms used by current network growth models.

3 PROBLEM STATEMENT

Extensive research on network growth has led to development of well-known growth models that generate realistic networks. However, the edge formation mechanisms of most network growth models tend to make strong assumptions about either knowledge (e.g. complete degree/fitness distribution known) or access (e.g. pick nodes uniformly at random).

The goal of this paper is to model network growth under information and resource constraints using edge formsation mechanisms. The growth model should be able to jointly explain global structural properties of real networks such as degree distribution, clustering coefficient distribution, degree-clustering relationship and degree correlations The model should incorporate information & resource constraints that influence edge formation in real networks.

4 CURRENT GROWTH MODELS

In this section, we describe four edge formation mechanisms underlying network growth models: preferential attachment [14] and its extensions, fitness [1], triangle closing mechanisms [5] and random walks [25]. These edge formation mechanisms explain the emergence of multiple structural properties of real networks, but make one or more strong assumptions.

Preferential attachment models explain the emergence of power law degree distributions of the form $p(k) = c \cdot k^{-\alpha}$, commonly observed in real networks. In preferential attachment models, new nodes that join the network form edges to existing nodes with probability proportional to their degree. This implies that high degree nodes accumulate edges quicker than low degree nodes. An intuitive explanation of preferential attachment is that new nodes are more likely to link to "popular" high degree nodes than relatively unknown, low degree nodes. However, preferential attachment implicitly assumes that edge formation depends only on degree and cannot explain why real networks exhibit high clustering or degree distributions that do follow power law.

Unlike preferential attachment models, fitness models are flexible enough to generate networks with varying degree distributions and degree correlation. The inability of preferential attachment to preserve multiple structural properties suggests that factors other than degree influence edge formation. In fitness models, new nodes that join the network form edges to existing nodes with probability proportional to their fitness. The fitness ϕ_i of node v_i is a function of intrinsic nodal properties that influence edge formation. The structural properties a fitness model preserves depends on the exact definition of fitness. For example, the fitness model introduced in [4] increases fitness as a function of degree and node recency. This can preserve temporal dynamics such as decay in popularity [27] of old nodes in citation networks. Simpler fitness models can generate degree distributions that follow power law, exponential or lognormal [21] distributions. However, since new nodes form each edge *independently*, fitness alone cannot explain the emergence of high local clustering or the bivariate relationship between degree and clustering observed in real networks.

Edge formation mechanisms proposed by the above network growth models make two strong assumptions.

- **Complete access to information** These mechanisms require nodes to link uniformly at random to *any* node in the network or have explicit knowledge of the degree/fitness of every node in the network. This assumption is unrealistic because nodes in real networks form edges partial information and limited access constraints.
- **Successive edge formations are independent** There is a strong, implicit assumption that a node's decision to link to another node is independent of the nodes to which it has already linked. This assumption contradicts a key empirical finding that the probability of edge formation increases as a function of neighborhood overlap [16] in social, information and citation networks.

Extensions of preferential attachment and fitness models [13, 15] using triangle closing mechanisms explain why social & information networks have high average local clustering [24]. In these models, a new node that joins the network "closes triangles" by linking to neighbors of nodes it has already linked to based on degree or fitness. Closing triangles increases the number of edges between neighbors, thereby increasing the average local clustering. Triangle closing mechanisms essentially model triadic closure, a sociological process that explains why two nodes with mutual neighbor(s) have an increased probability of connection. In Section 7, we show that these models are not flexible enough to capture the skewness and variance in clustering distributions of real networks.

A few models (e.g. [23]) adapt preferential attachment and fitness to model network growth under constraints of limited access and information. These models incorporate constraints by restricting access to recent nodes or a small set of nodes uniformly sampled from the network. However, these simple models are proof-of-concept methods that do not generate networks with varying degree distributions and realistic local clustering distributions.

Random walk models jointly explain multiple structural properties of real networks under constraints of limited access and information. New nodes explore neighborhoods of existing nodes without any assumption of global information and use simple rules to form edges. New nodes that join the network perform one or more random walks to link to existing nodes. For example, the Random Surfer model [6], in which new nodes link to the terminal nodes of short random walks, generate networks that exhibit power law degree distributions. Importantly, this model explains preferential attachment as an *emergent* property of local proceses. Random walk models in which new nodes perform random walk(s) and link to any visited node incorporate triadic closure and generate networks with heavy tailed degree distribution and high local clustering [12]. In Section 7, we show that models based on random walks outpeform well-known *global* edge formation mechanisms in preserving structural properties of citation networks. However, current random walk models are either inflexible or too simple to accurately capture local clustering observed in real networks.

To summarize, network growth models use one or more edge formation mechanisms to explain structural properties of real networks. Structural properties preserved by global edge formation mechanisms such as preferential attachment can be preserved by local processes such as random walks as well. However, unlike random walks, extensions of global processes such as preferential attachment & fitness models make strong, unrealistic assumptions.

5 EMPIRICAL ANALYSIS

In this section, we analyze the sturcture of citation networks to show that these networks exhibit similar structural properties. We begin by analyzing the rate of network growth and indegree distributions of citation networks. Then, we study the local clustering distribution and the relationship between indegree and local clustering. Finally, we briefly discuss the observed average path length. Figure 1 plots the structural properties of USSC and APS citation networks; The other three networks described in Section 2.2 have similar structural properties. Note that we preprocess the citation networks to remove a small fraction of nodes for which the time information is unknown. Finally, we conclude this section by motivating the need to study how the edge formation mechanisms that lead to these structural trends.

In many real networks, the average outdegree of nodes joining the network increases nonlinearly as a function of time and as a function of network size . Figure 1 shows that the average number of citations made by nodes drastically increases over time in both citation networks. Moreover, networks densify over times as the number of edges in the networks at time t, $e(t)$, increases superlinearly as a function of network size $n(t)$. Leskovec et al [19] show that densification in real networks exhibit a power law distribution $e(t) \propto n(t)^{\alpha}$ and can explain why the diameters of real networks shrink over time. Table 2 lists the densification power law (DPL) exponent α of all citation networks. In our proposed model, we increase the average outdegree of nodes that join the network to realistically model the rate at which real network grow.

Citation networks have highly skewed, heavy tailed indegree distributions. This suggests that while most nodes receive zero or a few citations, a small but significant fraction of nodes receive many citations and become "popular". This structural property is important because it helps test the extent to which popularity influences underlying edge formation mechanisms. Figure 1 shows the observed indegree distribution along with its power law fit for each citation network in blue and red respectively. While the power law fits can explain the heavy tail, it does not capture the initial concavity in the observed distribution. In Section 7, we show that our growth model can accurately capture the indegree distributions of citation networks in entirety.

The average local clustering coefficient (CC) in real networks tends to be high. Note that we define the neighborhood of node v as the set of nodes that cite v. Table 2 lists the average local clustering in all citation networks. High clustering indicates that a significant fraction of nodes that cite v tend to be connected to each other as well. Local clustering is fundamental to two well-known phenomena observed in real networks. First, clustering is one of the two components that explain the small-world phenomenon, in which two randomly picked nodes in large, sparse real networks are connected by a short path. Second, the clustering coefficient quantifies the extent to which triadic closure influences the underlying edge formation mechanism. By explicitly accounting for the fact that nodes are likely to link to neighbors of nodes it has already linked to, our model can not only generate networks with high average clustering but also capture the local clustering distribution observed in real networks.

We observe that the distribution over the local clustering coefficient of all nodes in real networks is skewed. Figure 1 depicts two local clustering distributions for each citation network; the observed distribution in solid blue and the distribution of a random network, generated using the observed degree sequence, in dashed green. The difference between the two distributions indicates that high local clustering is an inherent structural property of these networks. The skewness in these distributions highlights the high variance in the local clustering of real networks. Despite its widespread use, average local clustering coefficient is not a representative statistic of the skewed clustering distributions. As a result, network growth models that focus on generating networks with high average local clustering do not realistically capture the skewed clustering distribution of real networks. In Section 6 and Section 7, we show that our proposed edge formation mechanism can intuitively explain the emergence of the skewed clustering distribution observed in real networks.

In real networks, the average local clustering decreases as the indegree of a node increases [26]. This suggests that low indegree nodes have small, tightly knit neighborhoods and high indegree nodes have large, star-shaped neighborhoods. In Figure 1, we show that the degree-clustering relation in APS and USSC initially decreases as a linear function of the logarithmic value of indegree. In Section 7, we show that well-known growth models that generate networks with tunable average clustering are not flexible enough to explain the degree-clustering trend shown in Figure 1.

Citation networks are clustered, sparse networks that exhibit small average path length. Table 2 lists the average path length (APL) of all citation networks. We use a Monte Carlo method [18] to estimate the average path lengths as the citation networks are prohibitively large.

Table 2: Network characteristics for five real-world datasets. DPL: Densification power law, CC: clustering coefficient, APL: average path length

Network	DPL exponent	Avg. local CC	APL
HEP-PH	1.894	0.120	4.391
Patents	1.158	0.039	7.791
APS	1.334	0.108	5.001
Semantic	1.900	0.054	6.079
USSC	2.613	0.115	4.328

To summarize, citation networks are small-world networks that undergo accelerated network growth. These networks exhibit heavy tailed indegree distributions, skewed local clustering distributions and a negatively correlated degree-clustering relationship. The global structural similarity of citation networks prompts the question - do individuals use the same criteria to form edges?

In the next section, we propose a growth model that can jointly explain the emergence of these structural properties using a single edge formation mechanism

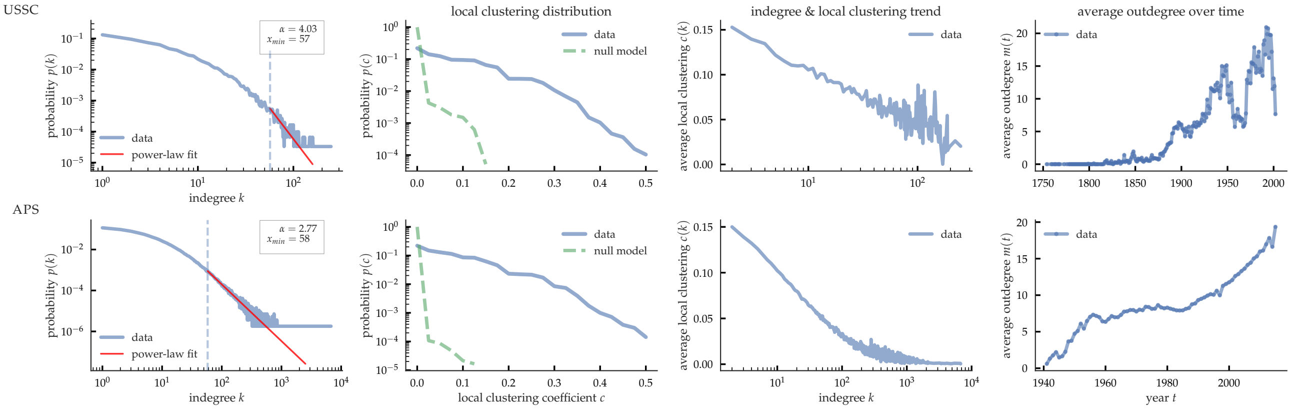

Figure 1: All citation networks exhibit similar network structural characteristics of heavy-tailed indegree distribution, skewed local clustering distribution, decreasing indegree & local clustering trend and increasing outdegree over time. We show the indegree distribution, clustering distribution, joint degree-clustering distribution and average out-degree over time for two representative networks—USSC.

6 PROPOSED MODEL

In this section, we present a resource-constrained growth model in which new nodes that join the directed network use a random walk process to link to existing nodes. In Section 6.1, we provide a detailed interpretation and description of our resource-constrained growth model. Next, in Section 6.2, we briefly explain the methods used to fit our model to observed networks. The goal of our resource-constrained model is to generate networks that follow structural properties of real networks discussed in Section 5.

6.1 Model Description

In this section, we describe three key components of our growth model. First, we explain how nodes join the network over time. Second, we describe how each node joins the network through an "entry point" under limited access constraint. Third, we describe the random walk mechanism that nodes use to form edges. We conclude by providing two natural interpretations of our growth model.

In our model, a directed network *grows* over time as new nodes join the network. The number of edges increases over time to reflect the nonlinear growth and densification of real networks [X]. More formally, at every discrete time step t, a new node u joins the network and forms m_u edges to existing nodes. At time $t = 0$, the initial network $\hat{G}_0$ consists of $|\hat{V}_0|$ nodes and $|\hat{E}_0|$ edges. Similarly, the network at time t, $\hat{G}_t$, consists of $|\hat{V}_t| = |\hat{V}_0| + t$ nodes and $|\hat{E}_t| = |\hat{E}_{t-1}| + m_u$ edges. In Section 6.2, we discuss the issue of initializing $\hat{G}_0$ and increasing the outdegree of new nodes over time.

The processes that new nodes use to select an entry point into the network and subsequently form edges intuitively corresponds to how we expect researchers to find references to cite. A researcher first finds one or more relevant paper as an "starting point". Then, under time and information constraints, he or she searches for potential references by navigating through a chain of references. After repeating this process one or more times, the researcher selects to cite a subset of these papers. Similarly, in our model, every node that joins the network selects a seed node from which it initiates the random walk process to search for potential links. Nodes terminate the random walk process after linking to a subset of visited node.

New nodes that join real networks select one or more "entry points" into the network under constraints of limited network access. We use a constant *recency* parameter $0 \leq p_r \leq 1$ to model the limited network access constraint under which nodes select entry points or seed node. Node u uniformly selects a seed node s_u from p_r fraction of nodes that have most recently joined the network. For example, if $p_r = 0.5$, a new node that joins the network at time t can only select a seed node that has joined the network after time $t/2$.

After selecting the seed node, a new node forms one or more edges to existing nodes. As discussed in Section 4 and Section 5, edge formation in real networks depend on local mechanisms such as triadic closure and do not require global information of every node in the network. In our model, new nodes use a random walk process to jointly explore the network and form edges. Random walks incorporates the idea of limited information and can only access its seed node and neighbors of nodes it visits. More formally, a new node u that joins the network at time t_u initiates a random walk from seed node s_u to form m_u edges.

The random walk process, visualized in Figure 2, can be described in four steps:

(1) At each step of the random walk, node u visits an existing node v_i. It links to this node with probability p_l.
(2) Then, with jumping probability p_j, u moves back to seed node s_u.
(3) Otherwise, with probability $(1 - p_j)$, u picks an outgoing edge with linking probability p_o *or* an incoming edge with probability $1 - p_o$, to visit a random neighbor of v. If v does not have any incoming edges, u picks an outgoing edge to visit a node neighboring v
(4) Node u repeats 1-3 until it forms m_u distinct edges.

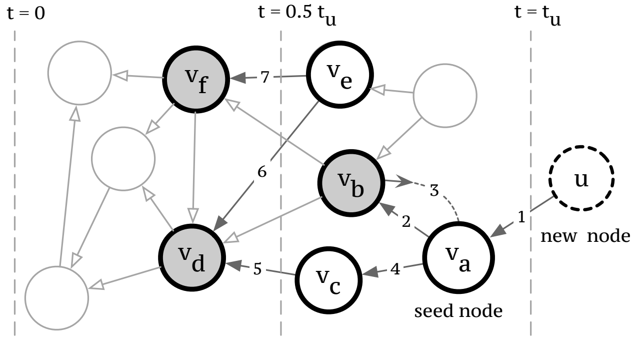

Figure 2: A toy example used for depicting the edge formation mechanism of the proposed random walk model. The recency parameter p_r of the random walker is 0.5. A new node u joins the network at time t_u with a prescribed outdegree of 3 and initiates a random walk from seed node v_a. The dark labeled edges denote the order in which node u traverses the graph using a random walk. Node u stops the random walk after linking to three nodes: v_b, v_d and v_f.

To summarize, we propose a growth model that incorporates constraints of limited network access and partial information that affect edge formation in real networks. In the next section, we show that our resource-constrained model can preserve key structural properties of real networks as well.

6.2 Model Fitting

Given a citation network $G = (V, E)$, a model fit should generate a directed network $\hat{G} = (\hat{V}, \hat{E})$ that preserves the structural properties observed in G. In this subsection, we describe methods to initialize $\hat{G}$, densify $\hat{G}$ over time and estimate the model parameters that generates networks structurally similar to G.

We now describe and justify the method used to initialize networks generated by our model. The random walk edge formation process is sensitive to a large number of weakly connected components in the initial graph $\hat{G}_0$. This is because a new node u that joins $\hat{G}$ cannot form edges to nodes that are not in the same weakly connected component as the seed node s_u. To ensure that the initial graph $\hat{G}_0$ is weakly connected, we perform an undirected breadth-first search on G starting from the oldest node that terminates after visiting 1% (0.1% if G large) of all nodes in G. The initial graph $\hat{G}_0$ is the small subgraph induced by the set of the visited nodes. After obtainining $\hat{G}_0$, new nodes sequentially join $\hat{G}$ and form edges using the random walk process until $|\hat{V}| = |V|$. (TODO: cold start problem)

Citation networks densify over time, with the number of edges growing superlinearly in the number of nodes. As shown in figure X, the average number of citations made by papers that join HEP-PH and APS per year increases in a nonlinear fashion. We incorporate densification into our model by increasing the outdegree of new nodes that sequentially join the network. Each new node u that joins the network $\hat{G}$ corresponds to some paper that joins the citation network G in year i. The number of edges that u forms is equal to the average number of the citations formed by papers that join G in year i. As a result, the rate of growth in networks generated by our model coarsely reflects the rate of growth in G.

The recency parameter p_r, link probability p_l, jump probability p_j and outgoing edge probability p_o jointly shape the random walk process that new nodes use to form edges. This subsequently determines the structural properties of the network $\hat{G}$ generated by the model. We use a straightforward grid search method to estimate the parameters values of p_r, p_l, p_j and p_o. In Section 7.1, we discuss the exact evaluation metrics and criteria used to select the parameter values that generate a network $\hat{G}$ most structurally similar to G.

To summarize this section, we first described and justified our growth model in which nodes use a random walk process to form edges under limited information and network access constaints. The growth model relies on four parameters: recency parameter p_r, link probability p_l, jump probability p_j and outgoing edge probability p_o. Then, we briefly discussed methods used to initialize $\hat{G}$, incorporate the observed growth rate into $\hat{G}$ and estimate the four model parameters. In the next section, we conduct experiments to evaluate whether our random walk model can jointly preserve structural properties of citation networks described in Section 2.2.

7 EXPERIMENTS

In this section, we present experimental results against four well-known baselines on citation networks described in Section 2.2. In Section 7.1, we describe the evaluation metrics and baselines used in our experiments. In Section 7.2, we describe and interpret the experimental results.

7.1 Experimental Setup

We first briefly summarize the baselines used in the experiments. Then, we describe the evaluation metrics used to quantify the extent to which growth models preserve structural properties of the citation networks.

We compare our model, abbreviated as RW, against four well-known *growth* models that are representative of the common edge formation mechanisms discussed in Section 4. Note that we do not consider graph generation models such as the Kronecker model [20] in which nodes do not join the network over time are not considered. The four baselines are:

- **Dorogovtsev-Mendes-Samukhin model** (DMS) [8] is an extension of the Barabasi-Albert model [X] that generates directed scale-free graphs using using preferential attachment. In this model, the probability of linking to a node is proportional to its indegree and "initial attractiveness".
- **Holme-Kim model** (HK) [13] is a preferential attachment model that generates scale-free, clustered, undirected networks using an additional triangle-closing mechanism. We modify the model to create directed edges and thereby generate directed networks.
- **Herera-Zufiria model** (HZ) [12] is a random walk model that generates scale-free undirected networks with "tunable" average clustering. We modify the model to generate directed networks by allowing the random walk process to traverse edge in any direction.

- **Forest Fire model** (FF) [19] is a recursive random walk model that generates directed networks which exhibit densification and decreasing diameter over time.

To ensure a fair comparison, we update the baseline models in two ways. First, models that do not have an explicitly defined initial graph use the initial network described in Section 6.2. Second, we extend models in which every node has the same outdegree to account for densification using the method described in Section 6.2.

Next, we describe the evaluation metrics used to measure the accuracy of the growth models in preserving the observed structural properties. We use three evaluation metrics in our experiments:

- **Kolmogorov-Smirnov (KS) statistic** computes the distance between univariate distributions such as indegree distribution & local clustering distribution of the observed network G and generated network $\hat{G}$.
- **Absolute difference** computes the distance between two point estimates such as the average local clustering.
- **Weighted relative error** measures the difference in the bivariate indegree-clustering trend of G and $\hat{G}$. The weighted absolute difference is defined as follows:

$$\sum_k p_G(k)\frac{|c(k)-\hat{c}(k)|}{c(k)}$$

The equation aggregates the weighted relative error between $c(k)$ and $\hat{c}(k)$, the average local clustering of nodes with degree k in networks G and $\hat{G}$ respectively. The weight of each term equals the probability mass of k in the indegree distribution $p_G(k)$ of the observed network G.

We estimate the four model parameters – recency parameter p_r, link probability p_l, jump probability p_j and outgoing edge probability p_o – using a grid search method to fit our model to a real network G. We select the model parameter values that minimize the L2 norm of the above evaluation metrics. We fit baseline growth models without a prespecified model fitting criteria using the same grid search method. After selecting the model parameters, our model can generate graphs that are structurally similar to the G. In the next section, we compare the perfomance of our model against the perfomance of four baseline models using the evaluation metrics discussed in this subsection.

7.2 Experimental Results

We present experimental results that demonstrate the effectiveness of our growth model in preserving three structural properties—indegree distribution, local clustering distribution and the indegree-clustering relationship—of the citation networks described in Section 2.2. We present the accuracy of our model and four baseline growth models in preserving the structural properties of the all five citation networks in Tables 3 to 5. Figure 3 illustrates the performance of all growth models in preserving the three structural properties of the APS network. We evaluate the performance of these structural properties using the evaluation metrics described in Section 7.1.

We now provide a brief overview of the experimental results followed by an interpretation of each result table. A common chararacteristic of the baseline growth models is that they cannot accurately preserve multiple structural properties observed in real networks. For example, the Dorogovtsev-Mendes-Samukhin (DMS) model can preserve indegree distribution but does not account for local clustering in real networks. Similarly, the Forest Fire model captures the skewed local clustering distribution in some networks but overestimates average local clustering as a function of indegree.

Table 3: Performance of baseline and proposed models (KS statistic) at modeling degree distribution. DMS is the best performing model for capturing degree distribution. It models the concavity and power-law of the degree distribution together. Our model RW is the second best model that outperforms all other competing baselines and does best on the largest dataset, Semantic Scholar.

	USSC	*HEP-PH*	*APS*	*Patents*	*Semantic*
DMS	**0.020**	**0.017**	**0.017**	**0.033**	0.052
HK	0.124	0.191	0.126	0.063	0.167
HZ	0.182	0.211	0.131	0.155	0.180
FF	0.168	0.171	0.277	0.141	0.121
RW	0.034	0.064	0.055	0.080	**0.025**

In table 3, we summarize the accuracy of each model in preserving the indegree distribution of citation networks. We observe that the DMS model performs better than our model RW by a small margin. This is because the model specifically captures the initial concavity in the distribution using an "attractiveness" parameter. Notice that the difference between DMS & RW and the other three models is significant.

Table 4: Performance of the baseline and proposed models (KS Statistic) at modeling the skewed clustering distribution of original network. RW models the clustering coefficient by tuning the parameter p_l and p_j that helps the random walker stay in the vicinity of seed node. This helps RW model the skewed clustering distribution while the other models fail to do so.

	USSC	*HEP-PH*	*APS*	*Patents*	*Semantic*
DMS	0.808	0.805	0.826	0.490	0.569
HK	0.415	0.480	0.525	0.062	0.147
HZ	0.087	0.273	0.338	0.081	0.090
FF	0.321	0.081	0.327	0.517	0.440
RW	**0.043**	**0.020**	**0.037**	**0.039**	**0.054**

Next, we show that the baseline growth models cannot accurately capture the skewness of the local clustering distribution in citation networks. Table 4 lists the KS statistic of each model for all citation networks. Our model outperforms the baselines by large margins as it captures the skewness of the observed clustering distribution in entirety. As shown in Figure 3, three out of four baselines – DMS, HK and HZ—do not capture the variance and skewness in the clustering distribution observed in the APS network.

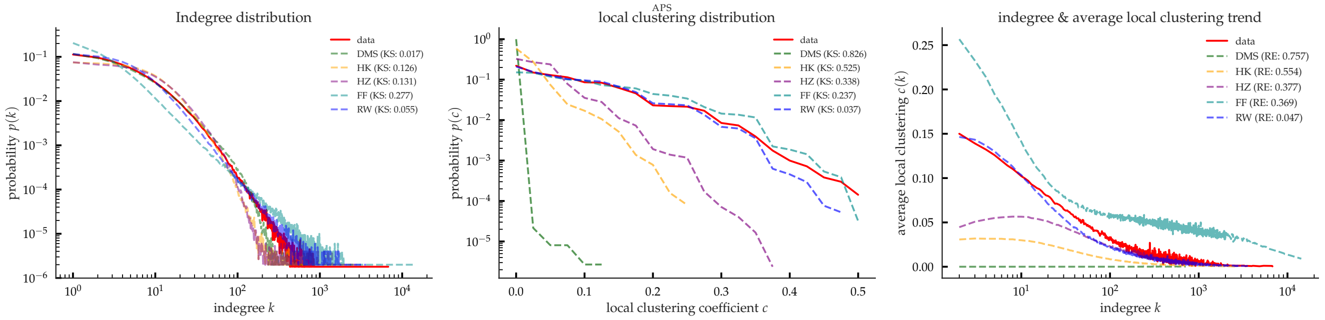

Figure 3: Accuracy of growth models at preserving structural properties of APS network. Our model RW outperforms the other in *jointly* preserving heavy-tailed indegree distribution, skewed local clustering distribution and the indegree & average local clustering trend.

Table 5: Performance of the baseline and proposed models (weighted relative error) at modeling the joint degree-clustering distribution of original network. RW model outperforms the baseline preferential attachment and random walk models by a signficant margin. The parameters p_o helps control the indegree and p_l & p_j helps control clustering of nodes by RW model. By simulatenously controlling both indegree and clustering, RW models the bivariate degree clustering trend with high accuracy.

	USSC	*HEP-PH*	*APS*	*Patents*	*Semantic*
DMS	0.657	0.681	0.757	0.589	0.592
HK	0.403	0.493	0.554	0.097	0.129
HZ	0.108	0.304	0.375	0.086	0.154
FF	0.437	0.504	0.369	2.023	1.170
RW	**0.038**	**0.052**	**0.047**	**0.048**	**0.086**

Next, we discuss the accuracy of the growth models in preserving average clustering as a function of indegree. Table 5 lists the weighted relative error, defined in Section 7.1, of each model for all citation networks. Our model outperforms the baselines by large margins. The DMS has the highest relative error as it does not preserve local clustering. As shown in Figure 3, the Holme-Kim HK and Herera-Zufiria HZ models that generate networks with tunable clustering underestimate the clustering of low-indegree nodes. Conversely, the Forest Fire (FF) model significantly overestimates the clustering of low-indegree nodes.

7.3 Parameter space of RW model

Through a series of extensive experiments, we observe that our model RW is able to model multiple structural characteristics of real-world networks. However, the fitted parameters are different for each dataset, suggesting possibly different local growth mechanisms in each network. Table 6 describes the best fitted parameters for five citation networks used in our experiments.

Table 6: Best fittedparameters obtained after grid search for random walk model.

	USSC	*HEP-PH*	*APS*	*Patents*	*Semantic*
p_l	0.80	0.80	0.15	0.25	0.40
p_j	0.30	0.65	0.65	0.05	0.15
p_o	0.95	0.95	0.80	1.00	0.95
p_r	0.50	0.80	0.85	0.45	0.60

To summarize, the experiment results on five citation networks against show that our resource-constrained model (RW) outperform four baseline growth models in accurately preserving degree, clustering and its relationship.

8 RELATED WORK

There has been extensive work on network growth models that *explain* how a subset of structural properties of real world networks emerge from edge formation mechanisms over time.

Preferential Attachment models such as the Barabasi-Albert model [3] and Vertex Copying model [17] show that power law degree distributions and small average path length in real networks can emerge as a result of preferential attachment. However, as shown in Section 7.2, models based on preferential attachment cannot explain properties related to local clustering.

As a result, richer edge formation mechanisms are needed to preserve multiple structural properties observed in real networks. Extensions of preferential attachment and fitness models account for clustering [13, 15] and degree correlation [7]. As described in Section 4, these extensions ignore the resource constraints that influence edge formation mechanisms and assume that successive edge attachments are independent.

Models such as [28, 29] adapt preferential attachment or fit to model network growth under more realistic constraints of partial information and/or limited access. However, such models are proof-of-concept methods that show that structural properties can arise from edge fomration under constraints. As a result, these models cannot accurately preserve multiple structural characteristics of real world networks.

Most similar to our work are random walk models that naturally incorporate resource constraints. Random walk models such as [6] in which nodes only form edges to terminal nodes of the walk preserve heavy-tailed degree distributions but cannot preserve local clustering. Random walk models [12] in which new nodes can form links to non-terminal nodse are flexible enough to generate networks with high average clustering. However, as shown in Section 7, these models do not necessarily capture the local clustering distributions or the degree-clustering properties observed in real networks.

In contrast, we propose a resource-constrained model that jointly preserves multiple structural properties using a single edge formation process. Unlike current random walk models, our model is flexible enough to capture the skewness and variance of the clustering distribution in real networks.

9 DISCUSSION

In this work, we address the problem of modeling growth of real-world bibliographic networks. Our proposed model is an improvement over existing random walk growth models that preserves multiple key network structural properties such as degree distribution, clustering coefficient distribution and their joint relationship. A standard modeling assumption is that *new* nodes joining a network can potentially make connection to *any* existing node in the network in some prescribed manner. Our experiments suggest that local link formation process in which *new* nodes explore local network neighborhoods and makes connections in the explored locality can explain multiple structural properties of real-world networks.

We note that clustering coefficient is an important characteristic of real-world networks. We observe that clustering is not uniformly distributed over the network and clustering at nodal level to be highly right skewed. The skewness implies that some parts of the network is more clustered than the other parts. In addition to skewness, clustering at nodal level is correlated to nodal degree. We propose a random-walk model the that gives rise to prominent characteristics of the network such as skewed local clustering.

10 LIMITATIONS

Now we discuss the limitation of our work. First, our work is limited to bibliographic datasets because of availibility of temporal data. We use the temporal out-degree sequence of incoming nodes in the network to model the network growth. In absence of temporal information, our growth model can be adapted by relying on the densification power law exponent. Second, our random walk model is sensitive to the initial graph. Since random walks explore the locality of a network and cannot access the entire network , the initial graph should have a giant weakly connected component. We recognise that the intialization problem can be addressed by having non-local source of information such as multiple seed nodes. Third, we note that our model fails to preserve certain network properties such as path length distribution and degree correlation. Limited by a completely local model, our model does not account for nodes that serve as “local bridges” in the network. As a result, it does not preserve the path length distribution. Modeling local and global process simulatenously in a joint random walk model should lead to preservation of the discussed key network properties.

11 CONCLUSION

In this paper, we model resource-constrained network growth model in which nodes use a random walk process to form edges under constraints of limited information and network access constraints. The problem is important because edge formation in real networks is usually a local process. In typical network growth scenarios, nodes in the network either have limited information about the other nodes in the network or the system allows access to only restricted portion of the existing network. It therefore becomes imperative to model how the local processes of link formation gives rise to network characteristics. In this work, we show that multiple structural properties of real networks can arise from the local process of exploration and link formation. Our results shows impressive performance over the next best competing model HZ [12] by a margin of 73.56% on the L2 norm of statistical difference in indegree distribution, local clustering and joint degree-clustering bivariate distribution of model and original network.

12 MODELING ATTRIBUTED NETWORKS

In this section, we extend our growth model to preserve properties of attributed networks. First, we briefly discuss attributed citation networks introduced in Section 2.2. Then, we describe an extension of our growth model that incoporates attributes into the random walk process. Finally, we present experimental results to show that our extended model can *jointly* capture both structure and attribute properties of real-world networks. To the best of our knowledge, there is no existing *growth* model that jointly and accurately preserves assortative mixing and *multiple* structural properties.

12.1 Attributed datasets

We now describe the attributed citation networks with their respective attributes.

Entities of citation network such as patents or papers have associated attributes which refer to as nodal attribute. For example, the PATENT dataset includes categories that classify patents based on academic discipline. Similarly, papers in the APS dataset are published at APS journals which can be treated as the attribute of a paper. In this work, we consider attributes that are inherent properties of the nodes and independent of structural features such as degree or clustering. Two nodes are similar if they share one or more attributes. Empiricial research [?] has shown that that similarity between two nodes can affect edge formation. Therefore, attributes influence the growth of networks and thus the global structural properties of networks.

Attributes in citation network often exhibit homophily or assortative mixing [?], which is the tendency of nodes to connect to similar nodes. Assortativity r is a commonly used metric to quantify the homphily (or heterophily) in a network. It is the difference between the fraction of edges between nodes of the same attribute value in the observed nework and the expected fraction of edges between nodes of the same attribute in a random null graph. Therefore higher assortativity values imply similar attribute nodes are more likely to form links than dis-similar attribute nodes. For example, the assortativity coefficient w.r.t. journal in APS is 0.459. This implies the papers in the same journal are have an increased likelihood of connection than papers in different journals.

12.2 Attributed random walk model

Now, we describe an extension of our growth model that can preserve assortative mixing in addition to the three structural properties: indegree distribution, local clustering distribution and indegree & local clustering trend. We refer to this model as the attributed growth model. The attributed growth model generates a directed network $\hat{G}$ with a single categorical attribute A. Similar to the original model, the attributed growth model has three components: node arrival, seed node selection and edge formation using random walks; we incorporate attributes in each component.

We make two assumptions about the attributes for modeling attributes of attributed networks. One, we assume that attributes are inherent properties of a node. Therefore, we use the original attribute distribution in datasets to model the attribute value of incoming nodes. Two, we assume that attribute-values have inherent ordering over one another. For example, journals often have a inherent "prestige" due to a variety of external reasons such as program committee, location and acceptance rate. Thus, we infer the "prestige" of an attribute using a affinity function $f(b)$ where $f(b)$ is the prestige of a journal b. We estimate the affinity function f from original dataset as the likelihood of the journal receiving a citation in the network.

Now, We state the formation of links by attributed random walk. We describe the modeling of attributed networks using just a single parameter p_s that controls the probability of a node to link to similar valued node.

- A new node u arrives the network. The node u draws it's out-degree m_u and attribute-value A_u from the underlying distribution.
- Node u makes a selection of seed node s_u. Node u first makes a decision for sampling seed node based on similarity with a probability of p_s and with a probability $1 - p_s$, node u samples the seed node based on its affinity score.
  - Under the first scenario, node u decides to draw the seed node based on similarity. Among p_r fraction of most recent nodes in the network, node u selects a similar valued node as the seed node.
  - Under the second scenario, node u decides to draw the seed node based on global affnity scores. It preferentially selects a seed node s_u from the network based on the seed node's affinity score which determined by affinity function f.
- After making seed node selection, node u performs a random walk as described in Section 6.1. While making connections, node u again flips a coin (with probability p_s) to form link either based on similarity or affinity as performed in the seed selection process.

The attributed random walk model is intuitive. The criteria used by new nodes to select seed nodes and form edges corresponds to how researchers find references to cite. For example, the assortative mixing w.r.t. paper journal in the APS network indicates the researchers tend to cite papers from the same journal. Moreover, papers published in "prestigious" or popular journals receive more citations than other journals, which indicates that researchers tend to cite papers in well-known journals. We model the effect of homophily and attribute-level affinity on edge formation by parameter p_s and affinity function f respectively.

12.3 Experimental results

Now, we show that our attributed growth model can jointly preserve attribute-based mixing patterns in addition to the structural properties preserved by the original model. We consider two attributed networks: APS and PATENTS that have attributes paper journal and patent category respectively. We select the model parameters using the grid search parameters described in Section 7.1.

As shown in **??**, our attributed growth model can preserve the assortativity w.r.t paper journal in the APS network up to three decimal places. The fixed similarity parameter p_s controls the preferences along the diagonal of the attribute mixing matrix and the affinity distribution f captures the popularity of journals PHYSREVB and PHYSREVLETT. In addition to assortativity, the growth model preserves degree distribution with KS statistic 0.036, local clustering distribution with KS statistic 0.026 and degree-clustering

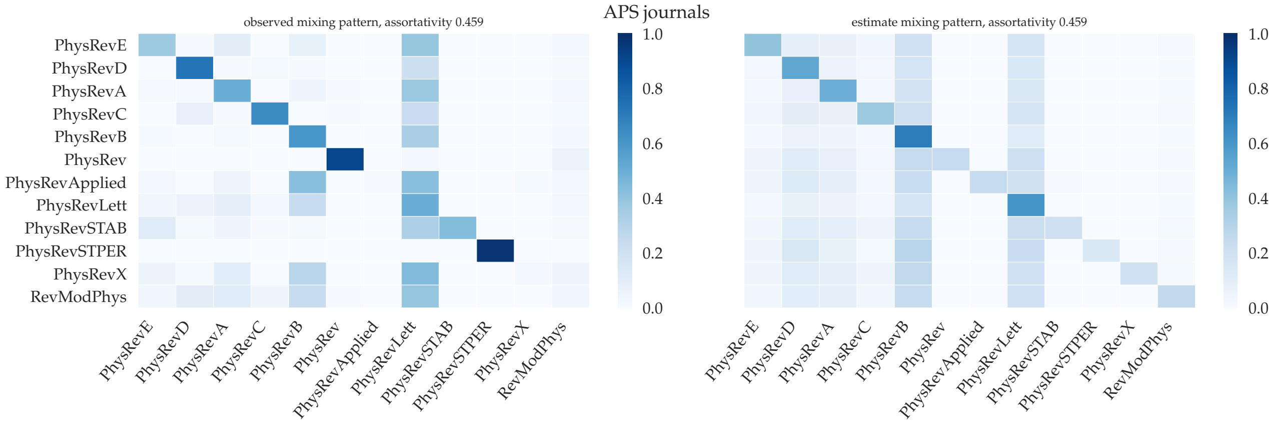

Figure 4: Attributed random walk model models the homophily in attributed networks with a very high accuracy ($r \approx$ 0.459 for both model and observed). Attributed random walker uses just one parameter p_s that controls the tendency of like valued nodes to link. This helps the random walker model not only homophily (assortativity) in the network but also other structural properties of the original network described in Section 7.1.

relationship with weighted relative error 0.108. Similarly for the PATENTS network, our growth model preserves degree distribution with KS statistic 0.10, local clustering distribution with KS statistic 0.036, degree-clustering relationship with weighted relative error 0.123 and assortativity with absolute difference 0.01.

To summarize, we show that our resource-constrained growth model can be extended to model the growth of attributed networks. We introduce two parameters—similarity p_s and affinity distribution f—to the growth model in order to jointly preserve attribute mixing patterns, indegree distribution, local clustering distribution and the indegree-clustering relationship.

REFERENCES

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[2] Albert-László Barabási. Linked: The new science of networks, 2003.

[3] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[4] Ginestra Bianconi and Albert-László Barabási. Bose-einstein condensation in complex networks. *Physical review letters*, 86(24):5632, 2001.

[5] Ginestra Bianconi, Richard K Darst, Jacopo Iacovacci, and Santo Fortunato. Triadic closure as a basic generating mechanism of communities in complex networks. *Physical Review E*, 90(4):042806, 2014.

[6] Avrim Blum, TH Hubert Chan, and Mugizi Robert Rwebangira. A random-surfer web-graph model. In *2006 Proceedings of the Third Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 238–246. SIAM, 2006.

[7] Michele Catanzaro, Guido Caldarelli, and Luciano Pietronero. Social network growth with assortative mixing. *Physica A: Statistical Mechanics and its Applications*, 338(1):119–124, 2004.

[8] Sergey N Dorogovtsev, Jose Ferreira F Mendes, and Alexander N Samukhin. Structure of growing networks: Exact solution of the barabási–albert's model. *arXiv preprint cond-mat/0004434*, 2000.

[9] Santo Fortunato, Marián Boguñá, Alessandro Flammini, and Filippo Menczer. Approximating pagerank from in-degree. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 59–71. Springer, 2006.

[10] James H Fowler and Sangick Jeon. The authority of supreme court precedent. *Social networks*, 30(1):16–30, 2008.

[11] Johannes Gehrke, Paul Ginsparg, and Jon Kleinberg. Overview of the 2003 kdd cup. *ACM SIGKDD Explorations Newsletter*, 5(2):149–151, 2003.

[12] Carlos Herrera and Pedro J Zufiria. Generating scale-free networks with adjustable clustering coefficient via random walks. In *Network Science Workshop (NSW), 2011 IEEE*, pages 167–172. IEEE, 2011.

[13] Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. *Physical review E*, 65(2):026107, 2002.

[14] Hawoong Jeong, Zoltan Néda, and Albert-László Barabási. Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, 61(4):567, 2003.

[15] Konstantin Klemm and Victor M Eguiluz. Highly clustered scale-free networks. *Physical Review E*, 65(3):036123, 2002.

[16] Gueorgi Kossinets and Duncan J Watts. Empirical analysis of an evolving social network. *science*, 311(5757):88–90, 2006.

[17] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D Sivakumar, Andrew Tomkins, and Eli Upfal. Stochastic models for the web graph. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 57–65. IEEE, 2000.

[18] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008.

[19] Jure Leskovec, Jon Kleinberg, and Christos 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.

[20] Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. Kronecker graphs: An approach to modeling networks. *Journal of Machine Learning Research*, 11(Feb):985–1042, 2010.

[21] Matúš Medo, Giulio Cimini, and Stanislao Gualdi. Temporal effects in the growth of networks. *Physical review letters*, 107(23):238701, 2011.

[22] Stanley Milgram and L van Gasteren. *Das Milgram-Experiment.* Rowohlt Reinbek, 1974.

[23] Stefano Mossa, Marc Barthelemy, H Eugene Stanley, and Luis A Nunes Amaral. Truncation of power law behavior in âĂIJscale-freeâĂİ network models due to information filtering. *Physical Review Letters*, 88(13):138701, 2002.

[24] Mark EJ Newman. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102, 2001.

[25] Alexei Vazquez. Knowing a network by walking on it: emergence of scaling. *arXiv preprint cond-mat/0006132*, 2000.

[26] Alexei Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67(5):056104, 2003.

[27] Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[28] Li-Na Wang, Jin-Li Guo, Han-Xin Yang, and Tao Zhou. Local preferential attachment model for hierarchical networks. *Physica A: Statistical Mechanics and its Applications*, 388(8):1713–1720, 2009.

[29] Jianyang Zeng, Wen-Jing Hsu, and Suiping Zhou. Construction of scale-free networks with partial information. *Lecture notes in computer science*, 3595:146, 2005.