Preferential Attachment in Graphs with Affinities

Jay-Yoon Lee* **Manzil Zaheer*** **Stephan Günnemann** **Alexander J. Smola**
School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213

Abstract

Preferential attachment models for random graphs are successful in capturing many characteristics of real networks such as power law behavior. At the same time they lack flexibility to take vertex to vertex affinities into account, a feature that is commonly used in many link recommendation algorithms.

We propose a random graph model based on both node attributes and preferential attachment. This approach overcomes the limitation of existing models on expressing vertex affinity and on reflecting properties of different subgraphs. We analytically prove that our model preserves the power law behavior in the degree distribution as expressed by natural graphs and we show that it satisfies the small world property. Experiments show that our model provides an excellent fit of many natural graph statistics and we provide an algorithm to infer the associated affinity function efficiently.

1 INTRODUCTION

Many natural systems can be represented as networks of interacting units: web pages connected through hyperlinks, molecular interaction in living cells, nerve cells in the brain, or the social network of interacting people are just a few examples. Modern network science explores graphs to identify the structural properties and come up with simple rules or generative processes capable of creating these networks.

Preferential attachment models have a long and prosperous history of modeling natural graphs. The simplest version is due to Erdős and Rényi [1] who propose a random link-generating process. Despite its simplicity this model proved to be very powerful in modeling important graph properties. That said, it fails to capture power-law behavior. The Buckley-Osthus [2] model addresses this by allowing for preferential attachment between vertices. That is, high-degree vertices are more likely to attract further links. Mimicking the properties of the Pitman-Yor [3] process allows one to obtain a large degree of attachment strategies.

At the same time, real-world link recommendation algorithms such as [4, 5, 6, 7] use a vast range of vertex attributes, in addition to the degree of a vertex, to suggest additional edges. This indicates that vertex to vertex affinity is quite unsurprisingly a key factor in social graphs. The same holds for inferring edges on the World Wide Web [8] — there the fact that two pages originate from the same domain or share a more detailed prefix is highly predictive for their propensity of establishing a link. Furthermore, it is desirable to model networks with heterophily towards particular attributes, such as gender, while showing homophily towards others, such as body shape, age or education, using the same framework.

The present paper addresses this issue by deriving an *integrated* model combining both preferential attachment and vertex affinity in a generative graph model. Our work borrows from a number of sources. Firstly, we build on the preferential attachment models of Buckley and Osthus [2] to address the power law behavior of vertex degrees. Secondly, we use affinity modifiers similar to the edge generation of the Distance Dependent Chinese Restaurant Process [9]. These allow us to model local variation in edge generation.

Outline: We begin reviewing related models in Sec. 2, laying the foundation for the Preferential Attachment in Graphs with Affinities, PAGA, in Sec. 3. In particular, we derive detailed properties such as the degree distribution, the diameter, and the clustering coefficient for PAGA graphs. Sec. 4 demonstrates that the derived properties closely match the observed statistics on real graphs. We present an affinity inference algorithm in Sec. 5, followed by a conclusion in Sec. 6.

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

*Both student authors had equal contribution.

2 RELATED WORK

Random graph generative models have a successful history, dating back to the works of Erdős and Rényi [1, 10, 11]. Their general approach begins by defining a probability space $\mathcal{G}_m^t$ over the set of all graphs having t vertices and an average degree of m per vertex. The goal is to study the properties satisfied by a typical random graph $G_m^t \in \mathcal{G}_m^t$. To be precise, we say G_m^t satisfies a property $\mathcal{P}$ if

$$\lim_{t\to\infty} \Pr(G_m^t \text{ satisfies } \mathcal{P}) = 1.$$

In other words we only require that in the limit of an infinitely large graph the property $\mathcal{P}$ holds with probability 1. One common way of obtaining graphs drawn from the Erdős-Rényi model is to draw each edge $(i \sim j)$ for $i, j \in G_m^t$ from a binomial distribution with probability $\frac{tm}{t(t-1)/2} = \frac{2m}{t-1}$.

While the Erdős-Rényi model has many desirable properties [12], it does not yield natural vertex degree distributions. Most importantly, it does not generate the ubiquitous power-law degree distributions [13, 14]. Accordingly, the question of graph generative models satisfying power laws has received much interest. Before going into further detail we summarize the used notation below.

V_t	Vertex set $V_t = \{1, 2, \ldots, t\}$
G_1^t	Graph on V_t with t edges
G_m^t	Graph on V_t with mt edges
$X(k,t)$	Set of vertices with degree k for G_1^t/G_m^t
$N(k,t)$	Expected size $X(k,t)$, i.e. $\mathbf{E}[X(k,t)]$
$\mathsf{d}_t(v)$	Degreee of node v in graph G_1^t
$\mathsf{d}_{t,i}(v)$	Degree of (super) node v in graph G_1^{mt+i-1}
$f(i,j)$	Similarity function between node i and j
$\bar{f}(\cdot,t)$	Average similarity between node t and nodes $\{1,\ldots,t-1\}$, i.e. $\frac{1}{t-1}\sum_{j=1}^{t-1} f(j,t)$

2.1 Bollobás-Riordian model

One of the most successful principles to realize power-law degree distributions are the preferential attachment random graph models, where new connections are made preferentially to more popular nodes. They were first introduced by [15], followed by a precise definition and analysis by Bollobás and Riordian [16].

The model proceeds in two stages: first an iterative process draws graphs G_1^t using preferential attachment akin to the Chinese Restaurant Process [17]: G_1^1 has a single vertex and a self loop. G_1^{t+1} is obtained from G_1^t by adding an edge $(t+1, v)$ according to

$$\Pr\{(t+1,v)\} = \begin{cases} \frac{\mathsf{d}_t(v)}{2t+1} & \text{for } v \le t \\ \frac{1}{2t+1} & \text{for } v = t+1. \end{cases} \quad (1)$$

By incorporating the degree of the vertices $\mathsf{d}_t(v)$, the preferential attachment is realized. To obtain graphs G_m^t with vertex degree higher than 1 one first generates G_1^{tm} and subsequently merges m sets of vertices $\{mi+1, \ldots m(i+1)\}$ into super nodes i in G_m^t. This construction yields to a degree distribution following a power law of the form k^{-3} [16].

2.2 Buckley-Osthus model

A problem of the above model is that the power-law exponent of 3 is by no means universal. This requires a correction of the preferential attachment rule (1) along the lines of the Pitman-Yor process [3]. Buckley and Osthus [2] modify (1) as follows:

$$\Pr\{(t+1,v)\} = \begin{cases} \frac{\mathsf{d}_t(v)+a-1}{(a+1)t+a} & \text{for } v \le t \\ \frac{a}{(a+1)t+a} & \text{for } v = t+1. \end{cases} \quad (2)$$

Here one uses a global attraction factor a to adjust the balance between the generation of self-loops and attachment to vertices with degree $\mathsf{d}_t(v)$. For $a \in \mathbb{N}$ and $0 \le k \le n^{\frac{a+1}{100}}$ one obtains graphs with power-law in degree distribution of k^{-2-a} [2].

Restrictions on the choice of k and the integrality constraint on a limit the applicability of the result. This was later generalized by [18], yielding a much for flexible power-law. Still, all of these models exploit only the structural information to generate the graph. In practice, however, vertex to vertex affinity based on further attributes is a key factor in social graphs.

2.3 Vertex attribute models

The above discussed models are successful in capturing the structural property of a network while lacking any information regarding vertex properties. We discuss some more common vertex attribute models below.

The Relational Topic Model of [19] takes both links and vertex properties into account. It builds on the Latent Dirichlet Allocation model of [20]: first all documents are generated using LDA. Subsequently links between pairs of documents are sampled according to a distribution that depends on the topics used to generate each of the constituent documents. Effectively this amounts to a Mixed Membership Stochastic Block Model [21] with an auxiliary document generating process. Such models efficiently yield joint models of side information and link structure.

Note, however, that such a process cannot generate power law degree distributions among the vertices, since it belongs to the family of infinitely exchangeable Bayesian graph models. The issue was pointed out by [22]. More specifically, [22] implies that any infinitely

jointly exchangeable graph must be either dense or empty. This can be seen as follows: by virtue of the Aldous-Hoover theorem [23] the adjacency matrix A of a graph must be drawn iid via

$$A_{ij} \stackrel{d}{=} F(U_i, U_j, U_{\{i,j\}}) \tag{3}$$

for a random function $F : [0,1]^3 \to 0,1$ where $(U_i)_{i\in\mathbb{N}}$ and $(U_{\{i,j\}})_{i,j\in\mathbb{N}}$ is a sequence and symmetric matrix of independent uniform $[0,1]$ random variables. This is equivalent to writing:

$$A_{ij} \stackrel{d}{=} \mathbb{I}_{\{U_{\{i,j\}} \leq W(U_i,U_j)\}} \tag{4}$$

for some function $W : [0,1]^2 \to [0,1]$. Thus, integration yields that $\Pr\{A_{ij} = 1\} = p$ for some $p \in [0,1]$. In other words, the number of edges is $\theta(t^2)$ for graphs with t vertices for any $p > 0$. Otherwise the graph is empty. The generative model is misspecified [22]. Similar degeneracies have been pointed out for the Exponential Random Graph Models (ERGM) yielding to undesired properties of the degree distribution [24, 25].

Recently, there have been efforts to overcome this limitation by moving from matrices/arrays to graphons which are functions over continuous domains [26, 27]. Although this approach seems to overcome the limitation of degeneracy, the practical implications of these models to real network data are still unclear [28].

As an alternative, our work addresses such problems by relying on a well-established graph generative model, which satisfies crucial graph properties, and we extend it by vertex affinities. This idea enables us to prove important properties rigorously and to reconstruct real graphs while preserving key statistics.

Further related models: The Multiplicative Attribute Graph (MAG) model [29] considers networks where each node is associated with a vector of categorical attributes. It requires a specific attribute-attribute similarity matrix to generate graph node affinities and it ignores the important aspect of preferential attachment. In contrast, our proposed PAGA model combines node similarities with node popularity and allows to incorporate any similarity function. Accordingly, as a further advantage, PAGA can also capture the correlations between features (cf. Sec. 5 for an example), while MAG is limited to independent features.

In [30], the principle of preferential attachment has been extended by assigning nodes to fixed states in an arbitrary metric space and by biasing the attachment probabilities according to an affinity function. Here, nodes have to be both similar and popular to realize linkage, leading to a rather restrictive model. PAGA, in contrast, considers a link probability model where nodes might be connected if they are either popular or similar. This principle leads to more realistic modeling of natural graphs, e.g., allowing to establish connections even if nodes are dissimilar (many websites link to Google for its popularity despite their dissimilarity). Furthermore, for PAGA we prove important characteristics such as the small world property and we conduct empirical studies to analyze whether the PAGA graphs match the statistics observed for real graphs.

PAGA combines preferential attachment with node affinities. As a different direction, [31] enhances the idea of preferential attachment by constraining linkage to nodes within a certain distance. Thus, intuitively, the similarity function between nodes corresponds to the indicator function based on the selected distance threshold. A generalization of this idea has been proposed in [32], leading to similar properties as mentioned for the work [30] above. While there is no discussion how to distribute nodes in the geometric space or how to specify reasonable distance thresholds to match properties of real graphs, in this paper we additionally provide a principle for learning the similarity function. Finally, in [31, 32] the power-law exponent of the degree distribution is limited to values larger than 3, while PAGA captures the properties of real graphs much better by realizing values larger than 2.

2.4 Distance dependent CRP

Related to our approach is the generative process used in the Distance Dependent Chinese Restaurant Process (ddCRP) of [9]. Note that the latter is only used to generate *partitions* of sets, while ignoring the edges generated in the process of inferring partitions.

In a nutshell, in the sequential version of the ddCRP nodes are added sequentially. A node $t+1$ is connected to another node v based on the following distribution:

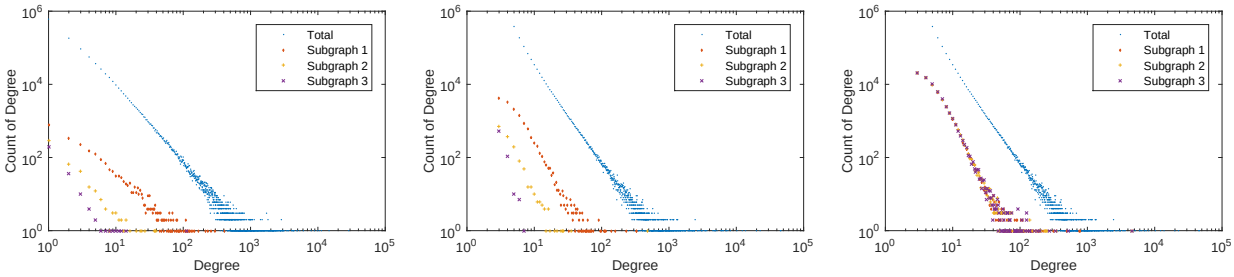
$$\Pr\{(t+1, v)\} = \begin{cases} \frac{f(v,t+1)}{\alpha + t\overline{f}(\cdot,t+1)} & \text{for } v \leq t \\ \frac{\alpha}{\alpha + t\overline{f}(\cdot,t+1)}, & \text{for } v = t+1 \end{cases} \tag{5}$$

Here f is an affinity function that decays with the 'distance' between vertices. See [9] for further details.

3 Preferential Attachment in Graphs with Affinities

3.1 Motivation and Contributions

While preferential attachment models successfully capture key properties such as the degree distribution, they fail to include intuitive reasons behind the link generation: the affinity between vertices based on their latent or observed attributes. It is natural that the generative process underlying the graph takes into account such vertex properties, e.g. web pages of similar content or within the same domain are much more likely to link to each other than dissimilar pages.



(a) Real graph (b) PAGA (c) B-O model

Figure 1: The real graph analysis in (a) reveals the fact that subgraphs follow distinct power laws. While PAGA can accomplish subgraph power laws (b), the B-O model fails to capture those properties (c).

As stated in its name, Preferential Attachment in Graphs with Affinities (PAGA) incorporates such vertex affinities as follows: while in the Buckley-Osthus model [33] the attraction a between vertices is constant, we vary a as a function $f(u, v)$ of the vertices (u, v) to be connected. This allows us to model local variations of attraction.

Using f not only allows to consider affinity but also enables to better represent the characteristics of community subgraphs: As observed in [34], the induced subgraphs of real communities in natural graphs show a power-law degree distribution and – even more important in our scenario – the distributions vary between different communities. To illustrate this more clearly, we analyzed in Figure 1a the YouTube dataset available at the Stanford Network Analysis Project (SNAP). This dataset (and further ones studied in [34]) contains externally provided community labels, representing real communities of the graph. Figure 1a shows the degree distributions of the induced subgraphs of three randomly selected communities as well as the degree distribution of the whole graph. Clearly, the power laws are different. As shown in Fig. 1b, PAGA can mimic this effect by using an appropriate affinity function. We will discuss the specific instantiation in more detail in Section 4. In contrast, the BO model cannot capture this effect, as evident in Fig. 1c.

Moreover, PAGA allows different inter and intra-community connections. To illustrate this, we drew a representative graph generated by the BO model for power law exponent 2.48 in Figure 2 (left column). This is compared to a graph generated by PAGA with same overall power-law of 2.48 (right column). In this scenario, the function f is instantiated such that 10 communities are generated, showing higher affinity between nodes within the same community (cf. Sec. 4). Clearly, community structure is evident as the subgraphs exhibit distinct connectivity. Yet, at the same time, both graphs exhibit essentially the same power law behavior (cf. lower part of Fig. 2). As shown, the PAGA model captures the same *global* characteristics as the BO model and additionally allows to model *local* characteristics of the graph based on vertex affinity.

In summary, our contributions are as follows:

- We give a construction for a random graph model which takes into account affinity among vertices.
- We obtain exact analytical answers for structural and topological characteristics of the graph.
- We compare properties of the graphs generated by PAGA to characteristics of real graphs.

As a supplementary contribution, we also discuss an inference algorithm to learn f based on observed data.

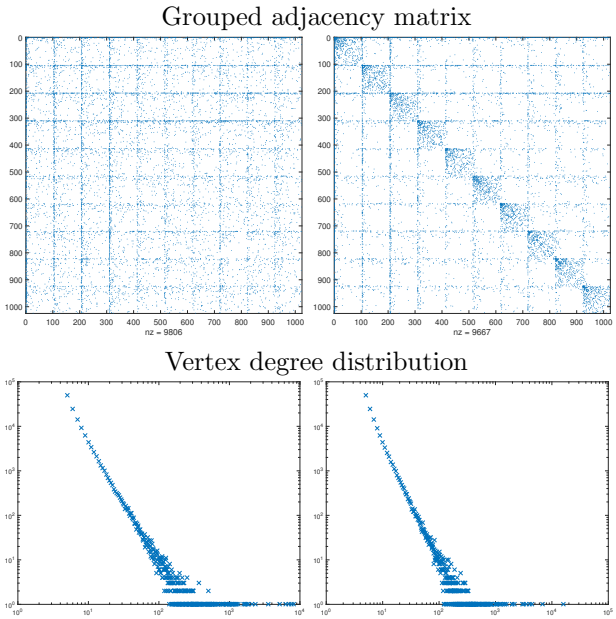


Figure 2: Left column: graph drawn using the Buckley-Osthus model. Right column: graph generated via PAGA. The BO model does not allow us to attach higher affinity within each subblock, as can be seen in the top row (in an unordered adjacency matrix both graphs would be visually indistinguishable). Note that both models exhibit the desired power law behavior.

3.2 Model

Our model combines ideas of the Buckley-Osthus model with the distance dependent Chinese Restaurant Process. For this purpose denote by $\mathcal{G}_m^t$ the following random graph process.

- Start with the undirected graph G_1^1 consisting of one vertex with a self-loop.
- Given G_1^t, PAGA models the probability of a new vertex $t+1$ to be connected with an existing vertex v via a single edge according to the following distribution:

$$\Pr\{(t+1,v)\} = \begin{cases} \frac{\mathsf{d}_t(v)+f(v,t+1)-1}{\left(\bar{f}(\cdot,t+1)+1\right)(t+1)-1} & \text{for } v \leq t \\ \frac{\bar{f}(\cdot,t+1)}{\left(\bar{f}(\cdot,t+1)+1\right)(t+1)-1} & \text{for } v = t+1 \end{cases}$$

  Here $\bar{f}(\cdot,t)$ denotes the average similarity between node t and nodes $\{1,\dots,t-1\}$, i.e.

$$\bar{f}(\cdot,t) := \frac{1}{t-1}\sum_{j=1}^{t-1} f(j,t).$$

  In other words, the new vertex $t+1$ connects to a random vertex v, where the probability that a vertex is chosen as v is proportional to its degree (at the current time) *as well as* its similarity to the new vertex.
- For the case $m > 1$, we add m edges from $t+1$ one at a time, counting the previous edges as well the edges being added as already contributing to the degrees. Equivalently, we collapse G_1^{mt} into G_m^t as in the Buckley-Osthus model.

It is clear from the construction that this model inherits the preferential attachment properties of the previos models. Moreover, f acts as a means for moderating the affinity between vertices. To verify our intuition, we now proceed to analyze important graph properties satisfied by the PAGA model.

3.3 Degree distribution

We will prove that the generated graphs show a power-law in degree distribution with exponent $2+\bar{f}(\cdot,t)$. Note that this new exponent has much higher degree of freedom compared to the previous random graph models as the term $\bar{f}(\cdot,t)$ depends on the similarities between node pairs.

Theorem 1 *Let $m \geq 1$ be a fixed integer. Assume that the similarity function satisfies $\forall u,v : f(u,v) \geq 0$ and $\Omega(\frac{1}{t}) \leq \overline{f}(\cdot,t) \leq O(1)$. Then in G_m^t, the expected value $N(k,t)$, for $k \geq m$ asymptotically follows a power-law, i.e. asymptotically we have $\frac{N(k,t)}{t} \sim k^{-2-\bar{f}(\cdot,t)}$. In fact we have a stronger result that:*

$$\begin{aligned} N(k,t) =& c(k,t,m+1)t + O_m\left(\frac{1}{k}\right) \\ c(k,t,i) =& \begin{cases} \frac{\mathrm{B}(k+m(\overline{f}(k,t,i)-1)),\overline{f}(\cdot,t)+2)}{\mathrm{B}(m(\overline{f}(m,t))),\overline{f}(\cdot,t)+1)}, & k \geq m, \\ 0, & k < m. \end{cases} \end{aligned} \tag{6}$$

where $c(k,t,m+1) \sim k^{-2-\bar{f}(\cdot,t)}$ asymptotically and $\overline{f}(k,t,i)$ is defined as:

$$\begin{aligned} \overline{f}(k,t,i) &= \frac{\sum_{j=1}^t \Pr(\mathsf{d}_{t,i}(j)=k) f(j,t+1)}{\sum_{j=1}^t \Pr(\mathsf{d}_{t,i}(j)=k)} \\ &= \frac{\sum_{j=1}^t \Pr(\mathsf{d}_{t,i}(j)=k) f(j,t+1)}{N(k,t,i)}. \end{aligned} \tag{7}$$

Here, $\mathrm{B}(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the Beta function which generalizes the combinatorial coefficient.

Proof: Our proof strategy is as follows[1]:

1. Express $N(k,t)$ in terms of $N(k,t,i)$, the expected value for the number of vertices of degree k not including the last node $(t+1)$ before the ith step.

$$N(k,t+1) = N(k,t,m+1) + \Pr(\mathsf{d}_{t,m+1}(t+1)=k) \tag{8}$$

   where $N(k,t,i+1)$ is defined as:

$$N(k,t,i+1) = \sum_{v=1}^{t} \Pr(\mathsf{d}_{t,i+1}(v)=k) \tag{9}$$

2. Show that

$$\Pr(\mathsf{d}_{t,m+1}(t+1)=k) = [k=m] + O_m\left(\frac{1}{t\bar{f}(\cdot,t+1)}\right) \tag{10}$$

3. Then it would suffice to show that $N(k,t,m+1)+[k=m] = c(k,t,m+1)t + O_m\left(\frac{1}{k}\right)$, by proving the following result:

$$\left| N(k,t,i) + \frac{i-1}{m}[k=m] - c(k,t,i)\left(t + \frac{i}{m} - \frac{1}{m(1+\overline{f}(\cdot,t))}\right)\right| = O_m\left(\frac{1}{k}\right) \tag{11}$$

4. We complete the proof by showing that $c(k,m+1)$ asymptotically behaves as $k^{-2-\overline{f}(\cdot,t)}$. □

To finish the study of the degree distribution, we show that $X(k,t)$ is concentrated around its expectation.

Corollary 2
As $t \to \infty$, $\frac{X(k,t)}{t} \to c(k,t,m+1)$ in probability.

[1]The full proofs of Theorem 1, 3, and 4 are available in the appendix.

Proof: Define a martingale process $Z_s = \mathbf{E}[X(k,t)|G_m^s]$ for $0 \le s \le t$. Note that $Z_t = X(k,t)$ while $Z_0 = \mathbf{E}[X(k,t)] = N(k,t)$ and $|Z_s - Z_{s-1}| \le 2m$ as two nodes which got connected at time s would not affect the degree distribution of other nodes. Using the Azuma-Hoeffding inequality [35, 36] we obtain:

$$\Pr\left(|X(k,t) - N(k,t)| > \epsilon\right) \le 2\exp\left(-\frac{\epsilon^2}{8t}\right)$$

Now by selecting $\epsilon = \sqrt{t\log t}$, we have:

$$\Pr\left(\left|\frac{X(k,t)}{t} - c(k,t,m+1)\right| > \sqrt{\frac{\log t}{t}}\right) \le 2t^{-1/8m^2} \to 0$$

as $t \to \infty$. □

3.4 Small world property and clustering coefficients

Next, we show that PAGA graphs satisfy the small world property, i.e. they show a small diameter.

Theorem 3 *Let $m \ge 1$ be a fixed integer. Assume that the similarity function satisfies $\forall u, v : f(u,v) \ge 0$ and $\Omega(\frac{1}{t}) \le \bar{f}(\cdot,t) \le O(1)$. Then in G_m^t, whp the diameter is at most $c\log t$, for some constant c.*

We conclude our analysis by studying the clustering coefficient.

Theorem 4 *If $\bar{f}(\cdot,t) \le 1$, then $C(G_m^t) \to 0$ for graph G_m^t generated from the PAGA model, where the clustering coefficient is defined as:*

$$C(G) = \frac{3 \times \textit{Number of triangles in } G}{\textit{Number of triplets in } G}$$

4 Experiments

In the following, we perform experiments to demonstrate the potential of PAGA to generate graphs matching the patterns occurring in natural networks. Please note that the aim of this section is not to derive the affinity function f best representing a given graph, but to empirically analyze the properties of PAGA graphs. While we discuss a principle to estimate f based on observed data in Section 5, in this part we focus on the following instantiation:

$$f(u,v) = [\pi \cdot \mathbf{B} \cdot \pi]_{u,v} \tag{12}$$

where π is a permutation matrix and $\mathbf{B}$ is a block-diagonal matrix with k homogenous blocks. That is,

$$\mathbf{B} := c_1 \cdot \mathbf{1}_{n_1} \oplus \ldots \oplus c_k \cdot \mathbf{1}_{n_k} \tag{13}$$

with $\oplus$ being the direct sum and $\mathbf{1}_n$ the all-ones matrix of size $n \times n$.

Effectively, this definition models a graph with k communities, each of size n_i, where the function $f(u,v)$ evaluates to c_i if both u and v belong to community i, and zero if they belong to different communities. Thus, for $c_i > 0$, higher affinity is obtained for nodes within the same community.

In our following experimental analysis (and likewise in Figure 2), we select $k = 10$, equally sized communities, and $c_i = 4.8$ for all i. Based on the definition of f, this leads to a value of $\bar{f}(\cdot,t) = 0.48$ and, thus, to a power-law exponent of 2.48 as already shown in Figure 2. By using non-equally sized communities and varying values of c_i, the behavior in Figure 1 can be obtained.

4.1 Empirical Analysis of Graph Properties

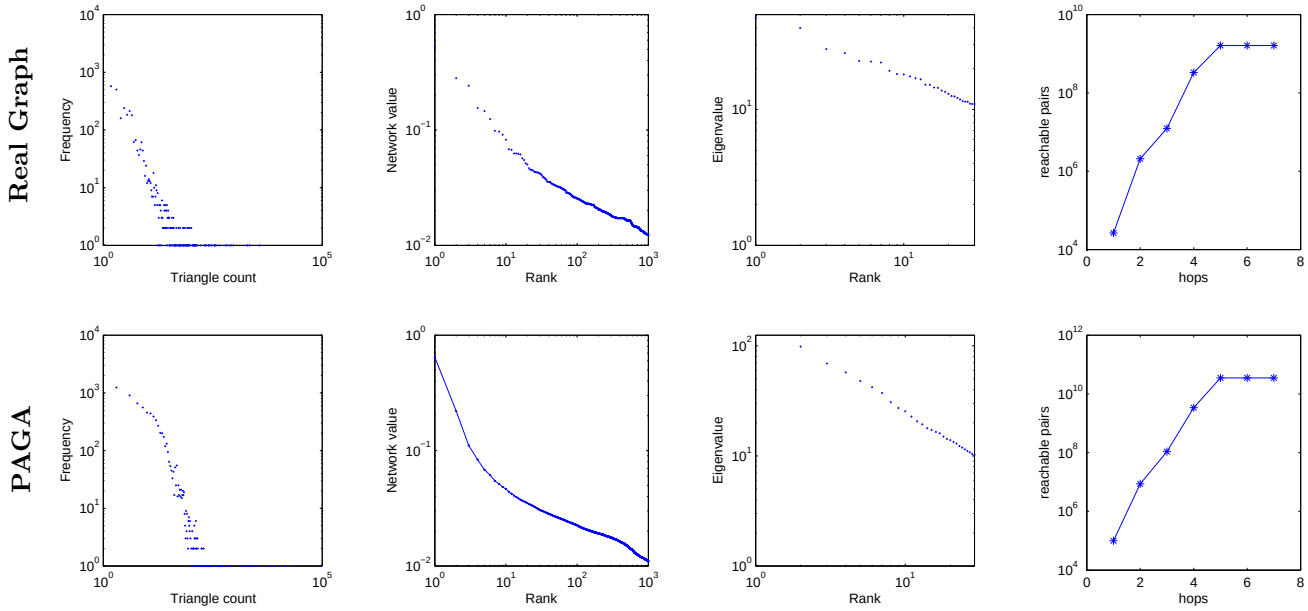
Degree distribution: The crucial characteristic of natural graphs is their power-law degree distribution, i.e. the number of nodes N_k with degree k is given by $N_k \propto k^{-\theta}$. Power law degree distributions have been observed in multiple domains [13, 38, 39] with exponents in the range 2.1 to 2.5. As proven in Theorem 1, PAGA ensures power-law distributions, with an example illustrated in Figure 2 (bottom right).

Besides the degree distribution, multiple other properties of natural graphs have been observed. In the following, we review the most prominent properties, we illustrate them exemplary on real networks, and we show that PAGA models similar properties. While Figure 3 analyzes properties referring to a single graph, Figure 4 compares statistics across multiple graphs.

Triangle counts: For natural graphs, it has been observed that the distribution representing the number of triangles x versus the number of vertices participating in x triangles is skewed [40].
Network value: The first/largest eigenvector of a graph, or more precisely the distribution of its components, is also known to be skewed [41].
Scree plot: The scree plot – illustrating the eigenvalues of the graph's adjacency matrix versus their rank in logarithmic scale – is also found to approximately follow a power law [41, 42].

Figure 3a-c illustrates these properties. As a representative, the top row shows the statistics derived from the *Autonomous systems* available at the Stanford Network Analysis Project (SNAP) [37]. The bottom row corresponds to PAGA. Clearly, the behavior matches.

Small Diameter: Small diameters (also known as the "small-world" phenomenon or "six degrees of separation" [43]) are ubiquitous in natural graphs. The diameter is the smallest number x such that every pair of



(a) Triangle distribution (b) Network plot (c) Scree plot (d) Hop neighbor

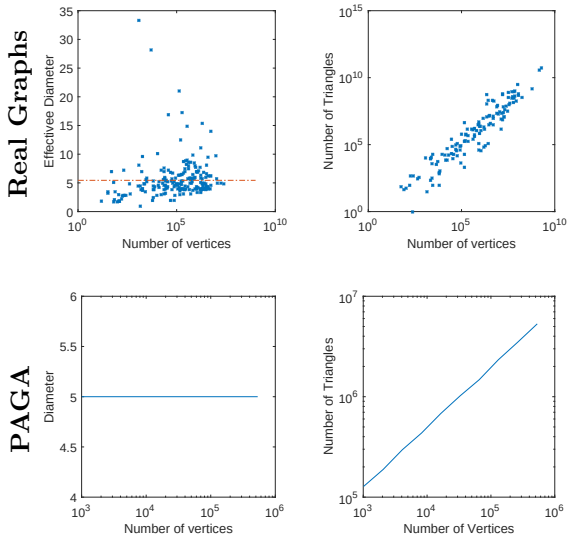
Figure 3: Properties of a real graph (top) and of PAGA (bottom). The real graph *Autonomous systems* [37] has 6474 nodes and 13895 edges. Without any fitting, PAGA reports mean statistics over 10 randomly generated graphs of 10k nodes and 50k edges with community affinities as stated in the beginning of Sec. 4. As shown, *Autonomous systems* and PAGA show similar skewed distributions on triangle, eigen values and eigenvector.

nodes can be connected by a path of at most x edges. Since the diameter is not robust to outliers, we refer to the effective diameter [44]. It corresponds to the minimum number of hops, in which a fraction (usually 90%) of all connected pairs of vertices can reach each other. Multiple studies [45, 43, 37] have observed small effective diameters in large natural graphs. Extending the idea of the diameter, a *hop plot* illustrates the number $g(h)$ of vertex pairs reachable within h hops [46].

The hop plot for a single graph is illustrated in Figure 3d, again showing the good fit of PAGA to the naturally occurring property. Similarly, we observe in Fig. 4 that the effective diameter across multiple graphs of different sizes is well represented by PAGA: In the figure, the top row shows statistics on more than 200 graphs from the KONECT website [47]. The red line in the top left figure indicates the average effective diameter of 5.1 across all graphs. In the bottom, the effective diameter of the PAGA graphs varying the size of the graphs is shown. With a value of 5, it is very close to the real graphs.

Finally, Fig. 4b confirms that also the overall number of triangles present in graphs generated by the PAGA model behave similar to natural graphs.

In summary, besides our analytical analysis, the empirical analysis supports that the PAGA model captures important properties of natural graphs. The preferential attachment strategy in combination with vertex affinity based on f realizes a flexible model. Next, we discuss how to estimate f based on observed data.



(a) Diameter (b) Number of Triangles

Figure 4: The diameter and number of triangles computed for graphs of different size. Top row: real graphs; bottom row: PAGA graphs.

5 Learning f

The affinity function f plays a crucial rule in our model. To mimic the behavior of real graphs, we will propose an efficient algorithm for learning f based on observed data. For this purpose, we assume a graph is given where nodes are enriched by vertex attributes, reflecting the various properties used, e.g., in link recommendation algorithms such as gender or age. Technically, we operate on the graph $G = (V, E, l)$ with vertices V, edges E, and labeling function $l : V \to \mathbb{R}^d$ mapping each vertex to a d-dimensional feature vector.

Our goal is to estimate f such that high affinity is obtained for adjacent nodes, while low affinity for non-adjacent nodes. This problem naturally casts to a binary prediction problem; consequently, we model it via a logistic regression approach, simultaneously ensuring efficient learning.

More precisely, we define f as $f(u, v) = \sigma(g(u, v))$ with σ being the sigmoid function and

$$g(u, v) = l(u)^T \cdot W \cdot l(v) \tag{14}$$

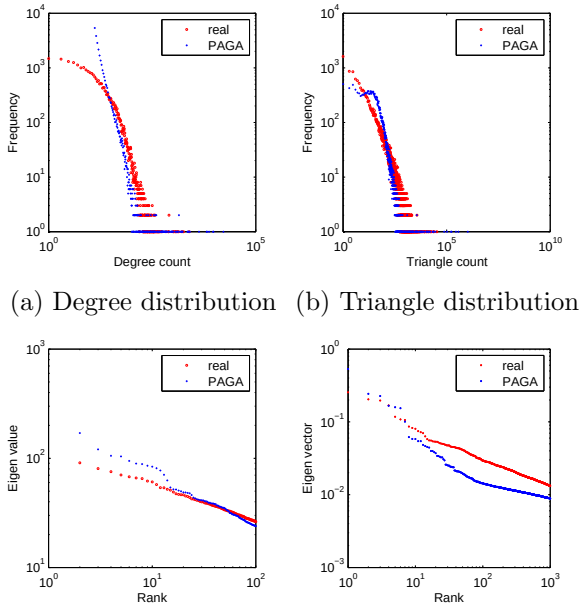
The matrix W captures the effects between the features' dimensions. In particular, in the case of (binary) feature vectors representing class memberships, one might interpret large diagonal entries in W as homophily within a class, while off-diagonal entries represent heterophily. Similarly, negative entries correspond to repulsion between different classes.

To learn W we refer to logistic regression, where the label for a pair of features is 1 if $(u, v) \in E$, and 0 otherwise. We add an L1 regularizer to enforce sparsity of the final matrix.

5.1 Case study on a citation network

In the following, we illustrate the potential of this learning method based on a document citation network (the high-energy physics theory citation network available from SNAP). We argue that the documents topics' are a leading factor for link establishment. Therefore, we exploit the documents' topic distributions as node features $l(u)$: we applied LDA, parametrized with 50 topics, on the papers' abstracts, leading to 50-dimensional feature vectors for each document.

Intuitively, since papers from the same topic more likely cite each other, we should observe high diagonal entries in W. A plot of W shown in the appendix confirms this hypothesis: most of the diagonal entries are positive with a high magnitude. Notably, some of the topics are less important for link establishment. In particular, topics covering very general terms (like topic 9 with "form, general, parameters, arbitrary") are less informative for the link establishment.



(a) Degree distribution (b) Triangle distribution

(c) Scree plot (d) Network plot

Figure 5: The red dots indicate a real citation network and the blue dots a PAGA graph fitted by learning the affinity function f. In (a), the real graph has a power-law degree distribution of 3.12 vs. the PAGA graph with 2.97. The total triangle count in (b) is 1.49m for the real graph and 1.44m for the fitted graph.

Given the learned function f, we show that realistic graph properties are obtained. Note that for this purpose we rescale the function f such that the value of $\bar{f}(\cdot, t)$ reflects the power-law exponent of the observed data. We then generate a graph according to PAGA, processing the documents and their corresponding feature vectors in the sequence of their publication time. As shown in Fig. 5, the properties of the real graph are very well matched by the generated counterpart.

6 Conclusion and Future Work

We proposed a random graph model combining the idea of preferential attachment and vertex affinity. By allowing local variations of the affinity between nodes, our model ensures more flexible power-law degree distributions and it captures local properties such a community structure. We analytically proved important structural properties of the generated graphs and we empirically showed that various further properties are well matched.

As future work we plan to develop an affinity learning technique handling very large graphs and we aim to estimate the sequence in which nodes and links have been added to an observed graph.

References

[1] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hungary. Acad. Sci.*, vol. 5, pp. 17–61, 1960.

[2] P. G. Buckley and D. Osthus, "Popularity based random graph models leading to a scale-free degree sequence," *Discrete Mathematics*, vol. 282, pp. 53–68, 2001.

[3] J. Pitman and M. Yor, "The two-parameter poisson-dirichlet distribution derived from a stable subordinator," *Annals of Probability*, vol. 25, no. 2, pp. 855–900, 1997.

[4] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks." 2006, pp. 611–617.

[5] S.-H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, and H. Zha, "Like like alike: joint friendship and interest propagation in social networks," in *International conference on World wide web*. ACM, 2011, pp. 537–546.

[6] M. De Choudhury, W. A. Mason, J. M. Hofman, and D. J. Watts, "Inferring relevant social networks from interpersonal communication," in *WWW*, 2010, pp. 301–310.

[7] F. Chierichetti, R. Kumar, S. Lattanzi, M. Mitzenmacher, A. Panconesi, and P. Raghavan, "On compressing social networks," in *ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 219–228.

[8] F. Chierichetti, R. Kumar, S. Lattanzi, A. Panconesi, and P. Raghavan, "Models for the compressible web," *SIAM Journal on Computing*, vol. 42, no. 5, pp. 1777–1802, 2013.

[9] D. M. Blei and P. I. Frazier, "Distance dependent chinese restaurant processes," *J. Mach. Learn. Res.*, vol. 12, pp. 2461–2488, 2011.

[10] P. Erdős and A. Rényi, "On random graphs," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[11] P. Erdős and A. Rényi, "On the strength of connectedness of a random graph," *Acta Mathematica Hungarica*, vol. 12, no. 1, pp. 261–267, 1961.

[12] R. Durrett, *Random graph dynamics*. Cambridge university press Cambridge, 2007, vol. 200, no. 7.

[13] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *SIGCOMM*, 1999, pp. 251–262.

[14] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[15] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[16] B. Bollobás, O. Riordan, J. Spencer, G. Tusnády *et al.*, "The degree sequence of a scale-free random graph process," *Random Structures & Algorithms*, vol. 18, no. 3, pp. 279–290, 2001.

[17] C. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Annals of Statistics*, vol. 2, pp. 1152–1174, 1974.

[18] E. A. Grechnikov, "Degree distribution and number of edges between nodes of given degrees in the buckley–osthus model of a random web graph," *Internet Mathematics*, vol. 8, no. 3, pp. 257–287, 2012.

[19] J. Chang and D. M. Blei, "Relational topic models for document networks," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 81–88.

[20] D. Blei and M. Jordan, "Modeling annotated data," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press, 2003, pp. 127–134.

[21] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed-membership stochastic blockmodels." vol. 9, pp. 1981–2014, 2008.

[22] P. Orbanz and D. M. Roy, "Bayesian models of graphs, arrays and other exchangeable random structures," *arXiv preprint arXiv:1312.7857*, 2013.

[23] O. Kallenberg, *Probabilistic symmetries and invariance principles*. Springer, 2005, vol. 9.

[24] T. Snijders, "Markov chain monte carlo estimation of exponential random graph models," *Journal of Social Structure*, vol. 3, no. 2, 2002.

[25] M. S. Handcock, G. Robins, T. A. Snijders, J. Moody, and J. Besag, "Assessing degeneracy in statistical models of social networks," Working paper, Tech. Rep., 2003.

[26] L. Lovász, *Large networks and graph limits*. American Mathematical Soc., 2012, vol. 60.

[27] P. J. Wolfe and S. C. Olhede, "Nonparametric graphon estimation," *ArXiv e-prints*, Sep. 2013.

[28] A. Z. Jacobs and A. Clauset, "A unified view of generative models for networks: models, methods, opportunities, and challenges," *ArXiv e-prints*, Nov. 2014.

[29] M. Kim and J. Leskovec, "Multiplicative attribute graph model of real-world networks," 1009.

[30] A. Santiago and R. Benito, "An extended formalism for preferential attachment in heterogeneous complex networks," *EPL (Europhysics Letters)*, vol. 82, no. 5, p. 58004, 2008.

[31] A. Flaxman, A. Frieze, and J. Vera, "A geometric preferential attachment model of networks," in *Algorithms and Models for the Web-Graph*, ser. Lecture Notes in Computer Science, S. Leonardi, Ed., vol. 3243. Springer Berlin Heidelberg, 2004, pp. 44–55.

[32] A. D. Flaxman, A. M. Frieze, and J. Vera, "A geometric preferential attachment model of networks ii," *Internet Math.*, vol. 4, no. 1, pp. 87–112, 2007.

[33] M. Zhukovskiy, D. Vinogradov, Y. Pritykin, L. Ostroumova, E. Grechnikov, G. Gusev, P. Serdyukov, and A. M. Raigorodskii, "Empirical validation of the buckley–osthus model for the web host graph: Degree and edge distributions," *ArXiv e-prints*, vol. abs/1208.2355, 2012.

[34] M. Araujo, S. Günnemann, G. Mateos, and C. Faloutsos, "Beyond blocks: Hyperbolic community detection," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I*, 2014, pp. 50–65.

[35] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

[36] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal*, vol. 19, no. 3, pp. 357–367, 1967.

[37] J. Leskovec, J. M. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24, 2005*, 2005, pp. 177–187.

[38] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "The web as a graph: Measurements, models, and methods," in *COCOON*, 1999, pp. 1–17.

[39] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener, "Graph structure in the web," *Computer Networks*, vol. 33, no. 1-6, pp. 309–320, 2000.

[40] C. E. Tsourakakis, "Fast counting of triangles in large real networks without counting: Algorithms and laws," in *Proceedings of the IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, 2008, pp. 608–617.

[41] D. Chakrabarti, Y. Zhan, and C. Faloutsos, "R-MAT: A recursive model for graph mining," in *Proceedings of the SIAM International Conference on Data Mining, Lake Buena Vista, Florida, USA, April 22-24, 2004*, 2004, pp. 442–446.

[42] I. J. Farkas, I. Derényi, A.-L. Barabási, and T. Vicsek, "Spectra of real-world graphs: Beyond the semicircle law," *Physical Review E*, vol. 64, no. 2, p. 026704, 2001.

[43] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[44] S. L. Tauro, C. R. Palmer, G. Siganos, and M. Faloutsos, "A simple conceptual model for the internet topology," in *Proceedings of the Global Telecommunications Conference, 2001. GLOBECOM '01, San Antonio, TX, USA, 25-29 November, 2001*, 2001, pp. 1667–1671.

[45] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[46] C. R. Palmer, P. B. Gibbons, and C. Faloutsos, "ANF: a fast and scalable tool for data mining in massive graphs," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, 2002, pp. 81–90.

[47] J. Kunegis, "KONECT: the koblenz network collection," in *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, 2013, pp. 1343–1350.

[48] B. Pittel, "Note on the heights of random recursive trees and random m-ary search trees," *Random Structures & Algorithms*, vol. 5, no. 2, pp. 337–347, 1994.

A Proof of Theorem 1

Here, we give a full proof of **Theorem 1**. Recall that an equivalent construction of the PAGA graphs is obtained by collapsing the graph G_1^{mt} to G_m^t leading to t super nodes. Accordingly, based on G_m^t, we can directly generate G_m^{t+1} by (temporarily) adding m more nodes, each linking with a single edge, and then collapsing these m nodes to the $(t+1)$th super node. Let $\mathsf{d}_{t,i}(v)$ denote the degree of the *super-node* v after $mt+i-1$ edges have been introduced, the probability that the $mt+i$th node (or equivalently its corresponding super-node $t+1$) connects to v is

$$\Pr\{(t+1,v)\} = \begin{cases} \frac{\mathsf{d}_{t,i}(v)+m(f(v,t+1)-1)}{(\overline{f}(\cdot,t+1)+1)(mt+i)-1} & 1 \le v \le t \\ \frac{\mathsf{d}_{t,i}(t+1)+i\cdot(\overline{f}(\cdot,t+1)-1)+1}{(\overline{f}(\cdot,t+1)+1)(mt+i)-1} & v = t+1. \end{cases} \tag{15}$$

Given these definitions, we begin with proving *statement 1*. With Eq. (9) and starting conditions

$$N(k,1) = [k = 2m], \quad N(k,t) = N(k,t,1) \tag{16}$$

we get a complete description of $N(k,t)$ as Eq. (8) of *statement 1*

$$N(k,t+1) = N(k,t,m+1) + \Pr(\mathsf{d}_{t,m+1}(t+1) = k)$$

Next we prove *statement 2* which is straightforward from the definition. First, note that $k \in \{m, m+1, ..., 2m\}$, i.e.:

$$\Pr(\mathsf{d}_{t,k+1}(t+1)k) = 0, \quad k < m \text{ or } k > 2m. \tag{17}$$

The minimal value $\mathsf{d}_{t,m+1}(t+1) = m$ is obtained when no one of the k edges is a loop. In this case, $\mathsf{d}_{t,i}(t+1) = i-1$ for all i, so:

$$\begin{aligned} \Pr(\mathsf{d}_{t,m+1}(t+1) = m) &= \prod_{i=1}^{m} \left(1 - \frac{i\cdot\overline{f}(\cdot,t+1)}{(\overline{f}(\cdot,t+1)+1)(mt+i)-1}\right) \\ &= 1 + O_m\left(\frac{1}{t\overline{f}(\cdot,t+1)}\right). \end{aligned}$$

From $\sum_{k=m}^{2m} \Pr(\mathsf{d}_{t,m+1} = k) = 1$ and $\Pr(\mathsf{d}_{t,m+1} = k) \ge 0$:

$$\Pr(\mathsf{d}_{t,m+1}(t+1) = k) = O_m\left(\frac{1}{t\overline{f}(\cdot,t+1)}\right), m < d \le 2m. \tag{18}$$

Finally the claim follows from the definition of $[k = m]$.

Before proving *statement 3*, we first derive some properties of $c(k,t,i)$ and prove *statement 4* en-route. Here, for brevity, we denote $c(k,t,i)$ as $c(k)$. Starting from $c(m)$ we study the step-wise change of $c(k)$ as:

$$\begin{aligned} c(m) &= \frac{B(m\cdot\overline{f}(m,t,i), \overline{f}(\cdot,t)+2)}{B(m\cdot\overline{f}(m,t,i), \overline{f}(\cdot,t)+1)} \\ &= \frac{\Gamma(\overline{f}(\cdot,t)+2)}{\Gamma(\overline{f}(\cdot,t)+1)} \times \frac{\Gamma(m\cdot\overline{f}(m,t,i)+\overline{f}(\cdot,t)+1)}{\Gamma(m\cdot\overline{f}(m,t,i)+\overline{f}(\cdot,t)+2)} \\ &= \frac{\overline{f}(\cdot,t)+1}{m\cdot\overline{f}(m,t,i)+\overline{f}(\cdot,t)+1} \end{aligned} \tag{19}$$

For $k > m$, the ratio of $c(k,i)$ to $c(k-1,i)$ can be simplified as:

$$\begin{aligned} \frac{c(k-1)}{c(k)} &= \frac{B\left(k-1+m(\overline{f}(k-1,t,i)-1)), \overline{f}(\cdot,t)+2\right)}{B\left(k+m(\overline{f}(k,t,i)-1)), \overline{f}(\cdot,t)+2\right)} \\ &= \frac{\Gamma\left(k-1+m(\overline{f}(k-1,t,i)-1)\right)}{\Gamma\left(k+m(\overline{f}(k,t,i)-1)\right)} \times \frac{\Gamma\left(k+m(\overline{f}(k,t,i)-1))+\overline{f}(\cdot,t)+2\right)}{\Gamma\left(k-1+m(\overline{f}(k-1,t,i)-1)+\overline{f}(\cdot,t)+2\right)} \\ &= \frac{k+m(\overline{f}(k,t)-1)+1+\overline{f}(\cdot,t)}{k-1+m(\overline{f}(k-1,t)-1)}, \end{aligned} \tag{20}$$

where we make the approximation $\overline{f}(k,t,i) \simeq \overline{f}(k-1,t,i)$ with the assumption that the average similarity of nodes with degree k to node t is similar to that of nodes with degree $k+1$. In particular, $c(k-1) > c(k)$, so $c(k) < c(m) < 1$ for all $k \geq m$.

Now to analyze the asymptotic behavior of $c(k)$ we begin by taking logarithm as:

$$\begin{aligned}\ln c(k,i) &= C_0 + \ln\Gamma(k+C_1) - \ln\Gamma(k+C_1+\overline{f}(\cdot,t)+2)\\ C_0 &= \ln\frac{\Gamma(\overline{f}(\cdot,t)+2)}{B(m\cdot\overline{f}(m,t,i),\overline{f}(\cdot,t)+1)}, C_1 = m(\overline{f}(k,t,i)-1)\end{aligned} \tag{21}$$

Exploiting the fact that asymptotically
$\Gamma(x+1) \sim \sqrt{2\pi x}\left(\frac{x}{e}\right)^x \xrightarrow{\ln} \ln\sqrt{2\pi} + x(\ln x - 1) + \frac{1}{2}\ln x$, Eq. (21) becomes:

$$\begin{aligned}&C_0 + (k+C_1)\left(\ln(k+C_1)-1\right) + O\left(\frac{1}{k}\right) - \left(k+C_1+\overline{f}(\cdot,t)+2\right)\left(\ln(k+C_1+\overline{f}(\cdot,t)+2)-1\right)\\ =&C_0 + (k+C_1)\left(\ln k + \frac{C_1}{k} - 1\right) + O\left(\frac{1}{k}\right) - \left(k+C_1+\overline{f}(\cdot,t)+2\right)\left(\ln k + \frac{C_1+\overline{f}(\cdot,t)+2}{k} - 1\right)\\ =&C_0 - \left(2+\overline{f}(\cdot,t)\right)\ln k + O\left(\frac{1}{k}\right)\end{aligned}$$

Rewriting the C_0 term, the full expression of asymptotic value of $c(k)$ as k grows becomes

$$c(k) = \frac{\Gamma(\overline{f}(\cdot,t)+2)}{B(m(\overline{f}(m,t,i)),\overline{f}(\cdot,t)+1)} k^{-2-\overline{f}(\cdot,t)}\left(1+O\left(\frac{1}{k}\right)\right).$$

To prove the *statement 3*, we show

$$\begin{aligned}&N(k,t,i+1)\\ &= \sum_{v=1}^{t}\Pr(\mathsf{d}_{t,i+1}(v)=k)\\ &= \sum_{v=1}^{t}(\Pr(\mathsf{d}_{t,i+1}(v)=k,\gamma=v) + \Pr(\mathsf{d}_{t,i+1}(v)=k,\gamma\neq v))\\ &= \sum_{v=1}^{t}(\Pr(\mathsf{d}_{t,i}(v)=k-1,\gamma=v) + \Pr(\mathsf{d}_{t,i}(v)=k,\gamma\neq v))\\ &= \sum_{v=1}^{t}\left(\Pr(\mathsf{d}_{t,i}(v)=k-1)\frac{k-1+m(f(v,t+1)-1)}{(\overline{f}(\cdot,t+1)+1)(mt+i)-1} + \Pr(\mathsf{d}_{t,i}(v)=k)\left(1-\frac{k+m(f(v,t+1)-1)}{(\overline{f}(\cdot,t+1)+1)(mt+i)-1}\right)\right)\\ &= N(k-1,t,i)\frac{k-1+m(\overline{f}(k-1,t+1,i)-1)}{(\overline{f}(\cdot,t+1)+1)\cdot(mt+i)-1} + N(k,t,i)\left(1-\frac{k+m(\overline{f}(k,t+1,i)-1)}{(\overline{f}(\cdot,t+1)+1)\cdot(mt+i)-1}\right)\end{aligned} \tag{22}$$

By proving *statement 4*, we complete our proof of **Theorem 1**. For the ease of proof, let us define $\tilde{N}(k,t,i)$:

$$\tilde{N}(k,t,i) = N(k,t,i) - c(k,t,i)\left(t + \frac{i}{m} - \frac{1}{m(1+\overline{f}(\cdot,t))}\right) \tag{23}$$

The lhs of (11) can be expressed in a recursive fashion as:

$$\begin{aligned}&\tilde{N}(k,t,i+1) + \frac{i}{m}[k=m]\\ &\quad= \tilde{N}(k,t,i+1) + \frac{1}{m}[k=m] + \frac{i-1}{m}[k=m]\\ &\quad= \tilde{N}(k,t,i+1) + [k=m]c(m,t,i)\frac{m(\overline{f}(m,t,i)) + \overline{f}(\cdot,t)+1}{m(1+\bar{f}(\cdot,t))} + \frac{i-1}{m}[k=m]\end{aligned}$$

$$
\begin{aligned}
&= N(k,t,i+1) - c(k,t,i)\frac{(\overline{f}(\cdot,t)+1)(mt+i+1)-1}{m(1+\overline{f}(\cdot,t))} - c(m,t,i)\frac{[k=m](m(\overline{f}(m,t,i))+\overline{f}(\cdot,t)+1)}{m(1+\overline{f}(\cdot,t))} + \frac{i-1}{m}[k=m] \\
&= N(k,t,i+1) - \left(\frac{(\overline{f}(\cdot,t)+1)(mt+i)-1}{m(\overline{f}(\cdot,t)+1)}\right) \\
&\quad \times c(k,t,i)\frac{(\overline{f}(\cdot,t)+1)(mt+i+1)-1-[k=m]\left(k+m(\overline{f}(k,t)-1)+\overline{f}(\cdot,t)+1\right)}{(1+\overline{f}(\cdot,t))\cdot(mt+i)-1} + \frac{i-1}{m}[k=m] \\
&= N(k,t,i+1) + \frac{i-1}{m}[k=m] - \left(t+\frac{i}{m}-\frac{1}{m(1+\overline{f}(\cdot,t))}\right)\frac{c(k,i)}{(1+\overline{f}(\cdot,t))(mt+i)-1} \\
&\quad \times \left\{(1-[k=m])\left(k+m(\overline{f}(k,t,i)-1)+1+\overline{f}(\cdot,t)\right) + \left((\overline{f}(\cdot,t)+1)(mt+i)-1\right) - \left(k+m(\overline{f}(k,t,i)-1)\right)\right\} \\
&= N(k,t,i+1) + \frac{i-1}{m}[k=m] - \left(t+\frac{i}{m}-\frac{1}{m(1+\overline{f}(\cdot,t)}\right) \\
&\quad \times \left(c(k-1,t,i)\frac{k-1+m(\overline{f}(k-1,t,i)-1)}{(1+\overline{f}(\cdot,t))\cdot(mt+i)-1} + c(k,i)\left(1-\frac{k+m(\overline{f}(k,t,i)-1)}{(1+\overline{f}(\cdot,t))(mt+i)-1}\right)\right) \\
&= \tilde{N}(k-1,t,i)\frac{k-1+m(\overline{f}(k-1,t,i)-1)}{(1+\overline{f}(\cdot,t))\cdot(mt+i)-1} + \tilde{N}(k,t,i)\left(1-\frac{k+m(\overline{f}(k,t,i)-1)}{(1+\overline{f}(\cdot,t))(mt+i)-1}\right) + \frac{i-1}{m}[k=m]
\end{aligned}
$$

Note that the last line comes from Eq. (22) and Eq. (23). Now, in a fashion similar to [18] involving manipulations like the ones done above, we have $\tilde{N}(k,t,i) + [k=m]\frac{i-1}{m} = O_m\left(\frac{1}{k}\right)$ finishing the proof of the theorem.

B Proof of Theorem 3

The proof of the bound on the diameter for PAGA graphs can by obtained by extending the original proof for the standard preferential attachment model from a uniform measure to a non-uniform measure, described by the affinity function f. Here, we ignore the case of having disconnected components in a graph as the graph will be one connected component with high probability.[2] As self loops do not affect the eccentricity of a node and hence the diameter, we simply ignore them in the generation process itself – for the purpose of bounding the diameter.

We begin the proof by noting that for $m = 1$ the graph is a tree and for the cases $m \geq 2$, which are formed by collapsing the graph G_1^{mt} graph, the diameter can only shrink. Next, the diameter of the graph can not be larger than twice the height of this tree, which is equal to the maximal graph distance between vertex 1 and any of the other vertices. So, it is sufficient to bound the height of the tree.

For bounding the tree height, we follow the steps of [48] and outline here the differences.[3] We start with a continuous time branching process, where the rate is given by $\lambda^t(j) = \mathsf{d}(j) + f(j,t)$. Therefore, the overall transition rate after t vertices are present (i.e. after $t-1$ births) is given by:

$$\sum_{j=1}^{t}(\mathsf{d}(j) + f(t,j) = 2t + t\bar{f}(\cdot,t+1) \tag{24}$$

Now we can decompose the time τ_t as a sum of independent variables, exponentially distributed with parameter $2t + t\bar{f}(\cdot,t+1)$, i.e.

$$\tau_t = \sum_{j=1}^{t} \mathfrak{t}_j$$

where $\mathfrak{t}_j \sim \mathrm{Exp}(2j + j\bar{f}(\cdot,j+1))$. It follows that the mean and variance are bounded by

$$\mathbf{E}[\tau_t] = \sum_{j=1}^{t}\frac{1}{2j+j\bar{f}(\cdot,j+1)} = O(\log t)$$

[2] In order for a PAGA graph to have a disconnected component, a new super-node $t+1$ has to make m self loops. This probability asymptotically goes to 0. Note that even in the uniform case this probability is $\frac{1}{t}^m$, while the self-loop probability in PAGA is typically smaller than $1/t$.

[3] To avoid clash of notations, we also redefine the birth times as $\tau_1, .., \tau_t$ instead of t_i.

$$\text{var}(\tau_t) = \sum_{j=1}^{t} \frac{1}{2j + j\bar{f}(\cdot, j+1)} = O(1)$$

These conditions match the ones required in [48]. Accordingly, the small world property holds for the PAGA model.

C Proof of Theorem 4

Note that at every timestep, there can be at most $\frac{m(m-1)}{2}$ new triangles added. So the number of triangles $T(G_m^t)$ is bounded by $O(n)$. On the other hand, the number of triplets in G_m^t follows $\sum_{k=1}^{d_{max}} N(k,t)\binom{k}{2} \propto \sum_{k=1}^{d_{max}} t \cdot k^{\bar{f}(\cdot,t)}$. As the sum $\sum_{k=1}^{d_{max}} k = mt$, it is straightforward that $C(G_m^t) \to 0$ unless $\bar{f}(\cdot,t) > 1$. We omit the analysis for the cases where $\bar{f}(\cdot,t) > 1$ as real graphs follow power law exponent around -2.1 to -2.5 ($0.1 \leq \bar{f}(\cdot,t) \leq 0.5$) □

D Spyplot of W

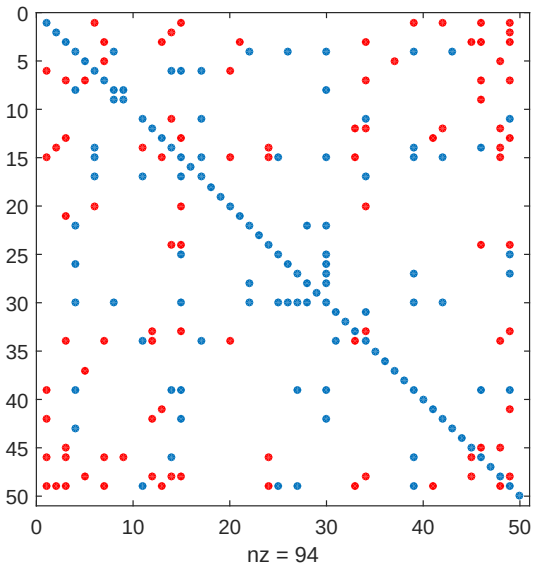


Figure 6: Spyplot of the matrix W. Blue entries correspond to positive elements; red entries to negative elements

E Topics and keywords learned by LDA

We used 50 topics (i.e. W is a 50×50 matrix) for learning the affinity function f. Here, we represent the top 20 keywords for top 10 topics we obtained from LDA.

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
method	quantization	gauge	theory	gauge
integral	hamiltonian	invariant	field	magnetic
path	dirac	invariance	approach	nonabelian
using	zero	local	based	abelian
functional	modes	brst	discuss	chernsimons
regularization	canonical	lagrangian	nonperturbative	electric
approach	physical	covariant	presented	charge
formula	lightcone	constraints	properties	theory
use	coordinates	formulation	new	monopole
integrals	shown	formalism	methods	su
expansion	quantized	class	may	monopoles
used	formalism	first	present	flux
series	variables	ghost	talk	charged
procedure	mode	cohomology	version	higgs
technique	constraint	extended	interpretation	dual
applied	leads	fields	developed	term
obtained	operator	lorentz	techniques	vortex
integration	approach	fixing	provide	selfdual
expression	constraints	transformations	used	vortices
formalism	formulation	auxiliary	analysis	yangmills
Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
n	matrix	coupling	loop	form
supersymmetric	model	point	theory	general
supersymmetry	matrices	limit	perturbation	parameters
supergravity	ansatz	fixed	renormalization	arbitrary
superconformal	spectral	large	wilson	given
super	models	strong	finite	explicit
superspace	relation	flow	perturbative	parameter
yangmills	limit	points	expansion	q
multiplet	integrable	constant	one	values
superfield	spin	group	order	set
multiplets	smatrix	renormalization	loops	independent
cal	scattering	weak	non	can
bps	elements	theory	result	terms
theories	eigenvalues	infrared	lambda	one
harmonic	chain	constants	cutoff	structure
susy	bethe	rg	diagrams	forms
superfields	corresponding	behavior	divergences	present
supersymmetries	factors	view	expectation	real
sym	related	critical	values	allows
large	representation	expansion	scheme	showns