Social-Affiliation Networks: Patterns and the SOAR Model

Dhivya Eswaran[1(✉)], Reihaneh Rabbany[2], Artur W. Dubrawski[1], and Christos Faloutsos[1]

[1] School of Computer Science, Carnegie Mellon University, Pittsburgh, USA
{deswaran,awd,christos}@cs.cmu.edu
[2] School of Computer Science, McGill University, Montreal, Canada
reihaneh.rabbany@mcgill.ca

Abstract. Given a social-affiliation network – a friendship graph where users have many, binary attributes e.g., check-ins, page likes or group memberships – what *rules* do its structural properties such as edge or triangle counts follow, *in relation to its attributes*? More challengingly, how can we synthetically generate networks which *provably* satisfy those rules or patterns? Our work attempts to answer these closely-related questions in the context of the increasingly prevalent social-affiliation graphs. Our contributions are two-fold: (a) **Patterns:** we discover three new rules (power laws) in the properties of *attribute-induced subgraphs*, substructures which connect the friendship structure to affiliations; (b) **Model:** we propose SOAR– short for SOcial-Affiliation graphs via Recursion– a stochastic model based on recursion and self-similarity, to *provably* generate graphs obeying the observed patterns. Experiments show that: (i) the discovered rules are useful in detecting deviations as anomalies and (ii) SOAR is fast and scales linearly with network size, producing graphs with millions of edges and attributes in only a few seconds. Code related to this paper is available at: www.github.com/dhivyaeswaran/soar.

Keywords: Graph mining · Attributes · Patterns · Anomalies Generator

1 Introduction

With the proliferation of the web and online social networks, social-affiliation networks – social/friendship networks where users have many, *binary* attributes or affiliations – have become increasingly common. Examples include social networking sites such as Facebook and Google+ which record user engagement, e.g., pages liked (attributes are pages – yes if liked, no if not), media-sharing social platforms such as Flickr and Youtube where users can form groups based on their interests (attributes are groups – yes if member, no if not), location-based social networks like GOWALLA where users can check-in at a location they physically visit (attributes are locations – yes if visited).

R. Rabbany—Work performed while at Carnegie Mellon University.

M. Berlingerio et al. (Eds.): ECML PKDD 2018, LNAI 11052, pp. 105–121, 2019.
https://doi.org/10.1007/978-3-030-10928-8_7

We consider two closely-related research questions concerning these networks: **[RQ1]** What *rules* (patterns) do the various structural properties of social-affiliation graphs – e.g., edge or triangle count – follow, *in relation to its attributes*? **[RQ2]** How can we synthetically generate realistic networks which *provably* satisfy these patterns? These questions fall under the umbrella of pattern analysis and modeling, a well-explored research area and a standard practice in understanding real-world graphs [6,16,17,19]. Our interest in considering these research questions stems in part from the scientific and practical impact that the works on pattern analysis and modeling have had in the past. The discoveries of the scale-free property (skewed degree distributions [10]) and the small world property (small graph diameters [28]) and respectively their preferential-attachment [4] and forest-fire [19] models, for instance, have had numerous applications in graph algorithm design, anomaly detection, graph sampling and more [3,18].

While works on patterns and models for non-attributed graphs abound in the literature, studies dealing with social-affiliation networks are somewhat limited [14,29] (see Sect. 2). Our work complements these by discovering rules which the structural properties of social-affiliation graphs follow *in relation to their attributes*. Specifically, we study "*attribute-induced subgraphs*" (AIS, in short) – each of which is a subgraph induced by the nodes affiliated to a given attribute – substructures which connect the structure of friendship graph to the distribution of attribute values. See Sect. 3 for more details and Fig. 1 for an example. Studying the patterns exhibited by the structural properties of AIS allows us to understand homophily effects ('birds of the same feather flock together') and consider questions of form 'If the number of users affiliated to attribute a doubles, what happens to the number of friendships between them?' As we show later, the patterns discovered based on AIS and the associated capability to answer 'what-if' questions are subsequently useful in (i) detecting anomalies and (ii) developing and testing a realistic model for social-affiliation graphs.

Our contributions are two-fold: **(a) Patterns:** We study four large real-world social-affiliation graphs and discover three new consistent patterns concerning the structural properties of attribute-induced subgraphs. With the help of a case study, we illustrate how the findings can be leveraged for anomaly detection. **(b) Model:** We propose the SOAR model to produce synthetic social-affiliation graphs *provably matching all observed patterns*. SOAR is based on self-similarity, implicitly incorporates attribute correlations, scales linearly with graph size and is up to 50× *faster* than the prior models for social-affiliation graphs.

Reproducibility. We use publicly-available datasets and open-source our code at www.github.com/dhivyaeswaran/soar.

2 Related Work

We group related work into three categories: models for social networks with no attributes [A] and those for social-affiliation graphs when attributes are given [B] and not given [C].

Table 1. Comparison with other models for social-affiliation graphs

Properties	MAG [15]	AGM [24]	Zhel [29]	SAN [14]	SOAR
Generates edges and attributes *simultaneously*			✔	✔	✔
Scalable with increasing number of edges and attributes			✔	✔	✔
Provably obeys all observed patterns					✔

[A] Social graphs with no attributes. Several outstanding network models have been proposed to explain the observed structural characteristics of real-world non-attributed networks. Notably, the Barabási-Albert model for heavy-tail degree distributions [4], Forest Fire model for shrinking diameter [19], Butterfly model for the evolution of giant connected component [20], Kronecker model for community structure [18] and Random Typing Graph Model for self-similar temporal evolution [2]. Excellent surveys are given in [6,13,22]. As such, it is not clear how these models could be extended to produce attributes, given the complex interplay between attributes and friendship structure [9,11,25].

[B] Social-affiliation graphs when attributes are given. The problem of modeling network structure in the presence of *known* nodal attributes has been studied. Notably, Multiplicative Attribute Graph (MAG) model [15] connects nodes according to user-specified attribute-based link affinities. Attributed Graph Model (AGM) [24] presents a generic approach using an accept-reject sampling framework to augment a given non-attributed network model with correlated attributes. Both MAG and AGM apply to settings with categorical (not just binary) nodal attributes; however, they scale poorly with the number of attributes: each edge is sampled proportional to roughly the dot product of nodal attribute vectors, which is an expensive operation, considering that the social-affiliation graph datasets we study have around $30K$ to $1.28M$ affiliations.

[C] Social-affiliation graphs when attributes are not given. The *simultaneous* generation of attributes and friendships, in the context of social-affiliation graphs (i.e., with many binary attributes), has received some attention. The pioneering work by [29] discovers several patterns in social-affiliation graphs (e.g., power law relation between number of friends and average count of affiliations). It proposes Zhel model by adapting the non-attributed microscopic graph evolutionary model [17] for this setting. [14] studies the evolution of directed social network of Google+ and its affiliations, focusing on the density, diameter, degrees and clustering coefficients of users and affiliations. It proposes SAN model augmenting [17] with attribute-augmented preferential attachment and triangle-closing mechanisms to replicate the observations on Google+. The patterns we discover in this paper are complementary to the above discoveries. Further, both Zhel and SAN model the *evolution* of social-affiliation graphs, by generating attributes and edges of one node at a time, while in contrast, we investigate a *one-shot* approach to graph generation (i.e., without modeling its evolution) which leads to *input parsimony and* $\sim$50$\times$ *speed-up* (see Sect. 5).

A qualitative comparison of social-affiliation graph models is given in Table 1.

Table 2. Frequently used symbols and their meanings

Symbol	Term	Description
$\mathcal{G}$	Social-affiliation graph	Undirected unweighted graph with many binary nodal attributes
n		Number of nodes in $\mathcal{G}$
k		Number of attributes in $\mathcal{G}$
$\mathbf{A}$	Adjacency matrix	$n \times n$ binary matrix showing edge existence
$\mathbf{F}$	Membership matrix	$n \times k$ binary matrix showing attribute possession
$\mathcal{G}_a$	Attribute-induced subgraph	Subgraph induced by nodes affiliated to attribute a
n_a	Node count	Number of nodes in $\mathcal{G}_a$
m_a	Edge count	Number of edges in $\mathcal{G}_a$
Δ_a	Triangle count	Number of triangles in $\mathcal{G}_a$
σ_a	Spectral norm	Highest singular value of the adjacency matrix of $\mathcal{G}_a$

3 Preliminaries

Notation. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{M})$ be a social-affiliation graph, where $\mathcal{V}$ is the set of nodes (users), $\mathcal{A}$ is the set of binary attributes (affiliations[1]), $\mathcal{E}$ is the set of unweighted undirected *who-is-friends-with-whom* edges among nodes and $\mathcal{M}$ is the set of *who-is-affiliated-to-what* attribute memberships between nodes and attributes. That is, if node u is connected to node u', then, $\mathcal{E}$ includes edges (u, u') and (u', u); similarly, $(u, a) \in \mathcal{M}$ iff node u is affiliated with attribute a. $\mathcal{G}$ is equivalently expressed as a tuple $(\mathbf{A}, \mathbf{F})$ of the $n \times n$ symmetric adjacency matrix $\mathbf{A}$ and the $n \times k$ membership matrix $\mathbf{F}$, where $n = |\mathcal{V}|$ and $k = |\mathcal{A}|$ denote the number of nodes and attributes respectively. The matrices are binary with 1 indicating the presence of an edge (in $\mathbf{A}$) or an attribute membership (in $\mathbf{F}$). Table 2 gives the frequently used notation.

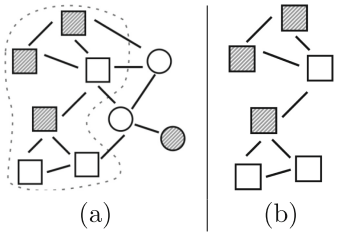


Fig. 1. (a) A social-affiliation graph with *isSquare*, *isStriped* attributes and (b) the subgraph induced by *isSquare* attribute

Attribute-Induced Subgraph (AIS). Given a social-affiliation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{M})$, the attribute-induced subgraph $\mathcal{G}_a$ corresponding to a given attribute $a \in \mathcal{A}$ is obtained by selecting the nodes affiliated to attribute a and the edges which link two such nodes. Formally, $\mathcal{G}_a = (\mathcal{V}_a, \mathcal{E}_a)$ where $\mathcal{V}_a = \{u \in \mathcal{V} \mid (u, a) \in \mathcal{M}\}$ and $\mathcal{E}_a = \{(u, u') \in \mathcal{E} \mid u, u' \in \mathcal{V}_a\}$. Let $n_a = |\mathcal{V}_a|$ and $m_a = |\mathcal{E}_a|$ denote its number of nodes and edges respectively. *Triangle count* Δ_a is the number of triangles in $\mathcal{G}_a$ while *spectral radius* σ_a is the largest eigenvalue of its adjacency matrix. An example of an AIS is given in Fig. 1.

[1] We use the following pairs of terms interchangeably throughout the paper: (graph, network), (node, user), (attribute, affiliation).

Datasets. We study four large publicly-available datasets, each of which contains a social network formed by friendship (or family) relations and also side-information regarding affiliations of users. Based on the nature of affiliations, we describe the datasets in two categories: *(i) Online-affiliation networks:* In FLICKR [21] and YOUTUBE [23], online photo-sharing and video-sharing websites respectively, users are allowed to form groups based on their common interests. We consider each group as a binary attribute, i.e., a user u has a group g if she participates in it. The friendship networks in these datasets are directed, but still, they have a high link symmetry or edge reciprocity [21]. Hence, for simplicity, we drop the direction of edges and retain a single copy of each resulting edge to get an undirected graph without multi-edges. *(ii) Offline-affiliation networks:* BRIGHTKITE and GOWALLA datasets [8] contain undirected friendship network along with user check-in information, i.e., who visited where and when. We use each location as a binary attribute; a user u has a location attribute l if she has visited l at least once. For a detailed description of these datasets, we refer readers to the papers cited above. Some useful statistics are provided in Table 3. The next section details our pattern discoveries on these datasets.

Table 3. Social-affiliation graph datasets studied

Dataset	Reference	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{M}\rvert$
YOUTUBE	[23]	$77K$	$0.4M$	$30K$	$0.3M$
FLICKR	[21]	$1.8M$	$16M$	$0.1M$	$8.5M$
BRIGHTKITE	[8]	$58K$	$0.2M$	$0.8M$	$1M$
GOWALLA	[8]	$0.2M$	$1M$	$1.28M$	$4M$

4 Pattern Discoveries

Given an attribute-induced subgraph $\mathcal{G}_a = (\mathcal{V}_a, \mathcal{E}_a)$, there is an infinite set of graph properties that one could investigate to look for patterns (number of nodes/edges, degree distributions, one or more eigenvalues, core number, etc.). Which ones should we focus on? Intuitively, we want to study properties that are (i) *fundamental*, easy to understand and interpret, (ii) *fast to compute*, exactly or approximately, in near-linear time in the number of edges and (iii) *lead to prevalent patterns* that AISs obey consistently across different datasets. After extensive experiments, we shortlist the following four properties of attribute-induced subgraphs: (i) $n_a = |\mathcal{V}_a|$: number of nodes in $\mathcal{G}_a$, i.e., number of users affiliated with attribute a. (ii) $m_a = |\mathcal{E}_a|$: number of edges in $\mathcal{G}_a$, i.e., number of friendships among users affiliated with attribute a. (iii) Δ_a: number of triangles in $\mathcal{G}_a$, typically indicative of the extent to which nodes in $\mathcal{G}_a$ tend to cluster together (e.g., via clustering coefficient). (iv) σ_a: spectral radius, or the principal eigenvalue of adjacency matrix of $\mathcal{G}_a$, roughly indicative of how large and how dense the giant connected component in $\mathcal{G}_a$ is. We list our observations regarding these properties in Sect. 4.1 and postpone explanations to Sect. 4.2.

4.1 Observations

Following standard terminology, we say that variables x and y obey a power law with exponent c, if $y \propto x^c$ [1]. Our pattern discoveries are all power laws with non-negative (and usually non-integer) exponents, as stated below.

Observation 1 ([P1] Edge count vs. node count). *Edge count m_a and node count n_a of AISs obey a power law: $m_a \propto n_a^{\alpha},\ 0 \leq \alpha \leq 2$.*

In the datasets we studied, $\alpha \in [1.17, 1.51]$. That is, double the nodes in an AIS, over double (roughly, triple) its edges.

Observation 2 ([P2] Triangle count vs. node count). *Triangle count Δ_a and node count n_a of AISs obey a power law: $\Delta_a \propto n_a^{\beta},\ 0 \leq \beta \leq 3$.*

In the datasets we studied, $\beta \in [1.24, 1.96]$. That is, as the number of nodes in an AIS doubles, its triangle count becomes about 3–4 times larger.

Observation 3 ([P3] Spectral radius vs. triangle count). *Spectral radius σ_a and triangle count Δ_a of AISs obey a power law: $\sigma_a \propto \Delta_a^{\gamma},\ \gamma \geq 0$.*

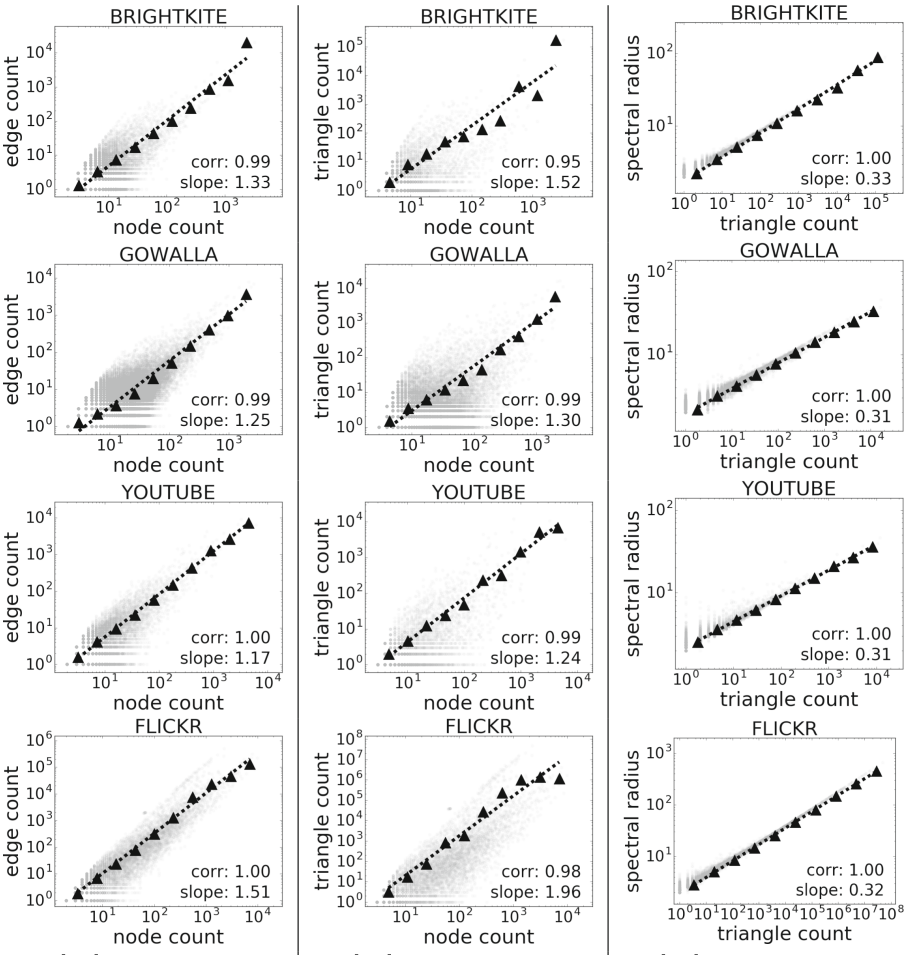
In the datasets we studied, $\gamma \in [0.31, 0.33]$. That is, doubling the spectral radius of an AIS leads to an eight-fold increase in its number of triangles.

Figure 2, which plots the relevant quantities (m_a vs. n_a, Δ_a vs. n_a and σ_a vs. Δ_a), illustrates these observations. The cloud of gray points in these figures show values corresponding to various AISs and darker areas signify regions of higher density. The relevant exponents α, β, γ are computed following standard practice (e.g., as in [16]). We bucketize x-axis logarithmically and compute per-bucket y averages (black triangles). The slope of the black line, which is the least-squares fit to the black triangles, gives the exponent. In addition, we report the *Pearson correlation coefficient* ρ of the per-bucket averages as a proxy for the goodness-of-fit of the power law relation. This value lies in $[0, 1]$ and intuitively, the higher the value is, the better is the fit. In our experiments, ρ was consistently above 0.95, suggesting a near-perfect fit.

4.2 Explanations, Use in Anomaly Detection, and Discussion

Here, we attempt to explain our observations in terms of known/expected properties of social-affiliation networks and hypothesize the nature of anomalies deviation from each pattern above would give rise to.

[P1] Edge count vs. node count. As the number of nodes in an AIS doubles, the number of edges remains the same ($\alpha = 0$) for empty social-affiliation graphs having no edges and quadruples ($\alpha = 2$) for complete graphs. As real-world social-affiliation networks tend to be sparse ($|\mathcal{E}| = \mathcal{O}(|\mathcal{V}|)$), one might expect the exponent α to be roughly 1. However, in experiments, α was *much higher*, e.g., ∼1.5 for Flickr dataset. This suggests *homophily*, i.e., more friendships among people sharing the same attributes, which causes the number of edges to more than double (in fact, triple) when the number of nodes is doubled.



[P1] Edge count vs. node count $\alpha \in [1.17, 1.51], \rho \geq 0.99$

[P2] Triangle count vs. node count $\beta \in [1.24, 1.96], \rho \geq 0.95$

[P3] Spectral radius vs. triangle count $\gamma \in [0.31, 0.33], \rho \geq 0.99$

Fig. 2. Patterns exhibited by attribute-induced subgraphs (each point is an AIS)

Attribute-induced graphs violating this pattern can be understood as *unusually sparse or dense* having too few/many friendships between users sharing an attribute, e.g., when no two people who go to Starbucks are friends with each other.

[P2] Triangle count vs. node count. As the number of nodes in an AIS doubles, triangle count remains the same ($\beta = 0$) for empty or tree/star-like graphs with no triangles and becomes eight times ($\beta = 3$) for fully connected graphs. In experiments, β was been 1 and 2; that is, the triangle count becomes 2–4 times when the node count doubles. This suggests that the AISs are nei-

ther stars nor cliques (as might ideally be expected based on homophily) but somewhere in between – consisting of several small stars, cliques and also possibly isolated nodes. Violations of this pattern can be understood as *unusually non-clustered* attribute-induced subgraphs (triangle-free, e.g., trees) or *unusually clustered* graphs (cliques). For example, it is suspicious if everyone who visits 'ShadySide' are friends with each other.

[P3] Spectral radius vs. triangle count. We know that the number of triangles in a graph is the sum of cubes of its adjacency's eigenvalues [12]. Based on this, we provide two sufficient conditions for the observed slope of $\gamma \approx 1/3$. *Condition 1 (Dominating first eigenvalue):* the first eigenvalue is much bigger than the rest; hence, triangle count of AISs are approximately the cube of their respective spectral radii (roughly, the number of triangles in their giant connected components, GCCs). *Condition 2 (Power law eigenvalues):* Lemma 1 provides an alternate explanation assuming exponents of eigenvalue power law distributions of all AISs are identical. Diving deeper into the eigenvalue vs. rank plots of AISs (see Fig. 3) reveals skewed eigenvalues distributions with similar slopes – suggesting that both reasons above are at play. Violations are due to attribute-induced subgraphs having *unusually small or sparse or dense GCCs*.

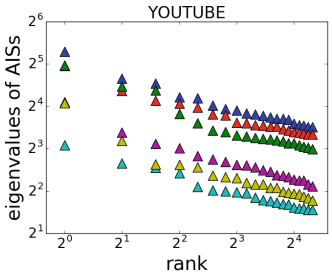


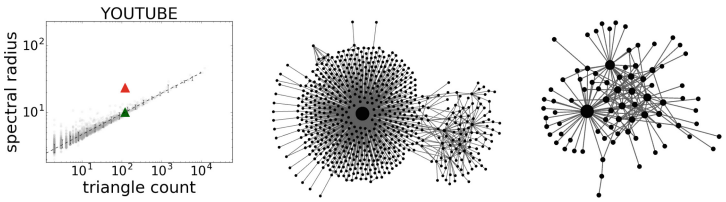
Fig. 3. Eigenvalues of 5 AISs with highest node counts from YouTube dataset

Lemma 1 (Spectral radius-triangle count power law). *If s is the common exponent of power law eigenvalue distributions of the attribute-induced subgraphs for a given social-affiliation graph, their triangle counts Δ_a and spectral radii σ_a approximately obey $\Delta_a = \sigma_a^3\ \zeta(3s)$ where $\zeta(\cdot)$ is the Riemann zeta function [27].*

Proof. As the eigenvalues of adjacency matrices of all AISs follow a power law with exponent s, the i^{th} eigenvalue of any AIS is $\sigma_a i^{-s}$, where σ_a is its spectral radius. Hence, triangle count Δ_a, which is the sum of cubes of eigenvalues of the adjacency, is equal to $\sum_i (\sigma_a i^{-s})^3 \approx \sigma_a^3 \sum_{i=0}^{\infty} i^{-3s} = \sigma_a^3\ \zeta(3s)$, as desired. □

Anomaly Detection. Our pattern discoveries represent normal behavior of attributes in a social-affiliation graph, deviations from which can be flagged as anomalies. For example, the spectral radius vs. triangle count plot for YouTube yields a dense cloud of points mostly distributed along a straight line in log-log scales (Fig. 4a); the red triangle marks an exception due to an anomalous attribute. It turns out that, as expected, the deviation was due to its *unusually sparse GCC*, which consisted of a giant star plus a few triangles (see Fig. 4b for its Gephi visualization [5]). In contrast, a typical AIS with a comparable triangle count (green triangle in Fig. 4a) has a denser GCC (Fig. 4c).

Discussion. It is natural to suppose that the data scraping methodology (sampling size/strategy) would have a considerable impact on the pattern discoveries.



(a) pattern (b) anomalous AIS ▲ (c) normal AIS ▲

Fig. 4. Anomaly detection using pattern [P3] reveals an attribute-induced subgraph (AIS) with an unusually sparse giant connected component (GCC) (Color figure online)

However, the consistency of our observations across datasets sampled in various ways – multiple sizes (GOWALLA and BRIGHTKITE – almost whole public data; FLICKR, YOUTUBE – large fraction of the giant weakly connected component [8,21]) and strategies (no sampling, snowball sampling using forward and/or reverse links depending on the public API) – suggest that the patterns are indeed generalize across many reasonable data scraping mechanisms. Also, note that our study is limited to the case of binary attributes; similar explorations of categorical and real-valued attributes are possible but left to future work.

5 SOAR Model

In this section, we show how to generate graphs which *provably* obey the discovered patterns using a *coupled* version of the matrix Kronecker product [26]. The resulting model, called SOAR– short for SOcial-Affiliation graphs via Recursion– has two steps: (i) an *initiator graph* $\mathcal{G}_1$, consisting of carefully coupled *initiator matrices* $\mathbf{A}_1$ for adjacency and $\mathbf{F}_1$ for membership, is chosen; (ii) the initiator graph is *recursively* multiplied with itself via *Coupled Kronecker Product* (Definition 2) for a desired number of steps to obtain the final social-affiliation graph. Sect. 5.1 presents SOAR model in detail. Our important contribution here is the proof that Coupled Kronecker Product is a *pattern-preserving* operation, i.e., if the initiator graph obeys patterns P1–P3, so does the final graph (see Sect. 5.2).

5.1 Proposed SOAR Model

Recall from Sect. 3 that $\mathcal{G}$ is a tuple $(\mathbf{A}, \mathbf{F})$ of the $n \times n$ symmetric adjacency matrix $\mathbf{A}$ and the $n \times k$ membership matrix $\mathbf{F}$, where $n = |\mathcal{V}|$ and $k = |\mathcal{A}|$ denote the number of nodes and attributes respectively. Given an initiator social-affiliation graph $\mathcal{G}_1 = (\mathbf{A}_1, \mathbf{F}_1)$, where $\mathbf{A}_1$ is the $n_1 \times n_1$ symmetric *initiator matrix* for adjacency and $\mathbf{F}_1$ is the $n_1 \times k_1$ initiator matrix for membership, we propose to derive the final social-affiliation graph $\mathcal{G} = (\mathbf{A}, \mathbf{F})$ via the recursive equation:

$$\mathcal{G}_{t+1} = \mathcal{G}_t \bar{\otimes} \mathcal{G}_1 \tag{1}$$

where $\bar{\otimes}$ is the Coupled Kronecker Product, as defined below:

Definition 2 (Coupled Kronecker Product (CKP)). *Given social-affiliation graphs* $\mathcal{G}_1 = (\mathbf{A}_1, \mathbf{F}_1)$ *and* $\mathcal{G}_2 = (\mathbf{A}_2, \mathbf{F}_2)$*, their Coupled Kronecker Product is given by*

$$\mathcal{G}_1 \bar{\otimes} \mathcal{G}_2 = (\mathbf{A}_1 \otimes \mathbf{A}_2, \mathbf{F}_1 \otimes \mathbf{F}_2) \tag{2}$$

where $\otimes$ *is the matrix Kronecker product.*

After M steps of Eq. (1), we obtain a $n \times n$-dim $\mathbf{A}_M$ and a $n \times k$-dim $\mathbf{F}_M$ where $n = n_1^M$ and $k = k_1^M$ respectively. When the initiator matrices are binary, so are the final matrices and thus can be directly used as the adjacency $\mathbf{A}$ and membership $\mathbf{F}$ matrices, respectively. It turns out that the above process captures the required power laws but has several discrete jumps (fluctuations). Hence, we use the stochastic version below.

The main idea is to produce at every recursive step, matrices of edge/membership occurrence probabilities instead of discrete (binary) edges/memberships. Thus, we begin with initiator matrices having real number entries in $[0, 1]$ (they do not need to sum to 1) and add a small relative noise η to the initiator matrices independently at every recursive step t. This process results in the final dense probability matrices $\mathbf{A}_M$ and $\mathbf{F}_M$, from which we recover $\mathbf{A}$ and $\mathbf{F}$ by sampling each entry proportional to its final value. A scalable implementation of the above approach by sampling one edge or membership at a time is given in Algorithm 1. The Hadamard product $\odot$ in lines 6 and 8 performs an element-wise matrix multiplication to add the desired noise to the initiators.

Running Time Analysis. Initialization (*ln 1–11*) contributes a fixed overhead of $\mathcal{O}(M(n_1^2 + n_1 k_1))$. The generation of edges (*ln 12–20*) and memberships (*ln 21–29*) take $\mathcal{O}\left(n_1^2 M\right)$ per edge and $\mathcal{O}\left(n_1 k_1 M\right)$ per membership respectively. As n_1, k_1 and M are small in practice (<10), Algorithm 1 is *linear in the number of edges and attribute memberships.*

5.2 Theoretical Properties

The structural properties of graphs generated using Kronecker product are well-studied and a number of desirable properties have been proved, e.g., multinomial distribution of degrees and singular values, etc. [18]. These properties directly carry over to the proposed model. More surprisingly, for careful coupling of initiators, SOAR graphs provably obey all the discovered power laws from Sect. 4. This is due to the *pattern-preserving* property of the Coupled Kronecker Product operation. That is, if graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ obey the patterns P1–P3 with the same exponent, then, so does their Coupled Kronecker Product $\mathcal{G}_1 \bar{\otimes} \mathcal{G}_2$. This is stated in Lemmas 3–5 (proofs in appendix).

Lemma 3 (CKP preserves [P1]). *If* $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *obey the edge count vs. node count power law with exponent* α*, i.e.,* $m_a \propto n_a^{\alpha}$*, so does* $\mathcal{G}_1 \bar{\otimes} \mathcal{G}_2$*.*

Lemma 4 (CKP preserves [P2]). *If* $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *obey the triangle count vs. node count power law with exponent* β*, i.e.,* $\Delta_a \propto n_a^{\beta}$*, so does* $\mathcal{G}_1 \bar{\otimes} \mathcal{G}_2$*.*

Algorithm 1. SOAR model

input : $\mathbf{A}_1 \in [0,1]^{n_1 \times n_1}$, $\mathbf{F}_1 \in [0,1]^{n_1 \times k_1}$, $M \in \mathbb{N}$, $\eta \in [0,1]$
output: $(\mathbf{A}, \mathbf{F}) = \text{SOAR}(\mathbf{A}_1, \mathbf{F}_1, M, \eta)$

num_edges $\leftarrow \lfloor (\text{sum of entries in } \mathbf{A}_1)^M \rfloor$
num_memberships $\leftarrow \lfloor (\text{sum of entries in } \mathbf{F}_1)^M \rfloor$
/* create M noisy copies of initiators $(\mathbf{A}_1, \mathbf{F}_1), \ldots, (\mathbf{A}_M, \mathbf{F}_M)$ */
$\mathbf{A}_0, \mathbf{F}_0 \leftarrow \mathbf{A}_1, \mathbf{F}_1$
for $t = 1, 2, \ldots, M$ **do**
  Sample $\mathbf{N}_{\mathbf{A},t} \sim [-0.5, 0.5]^{n_1 \times n_1}$ // i.i.d, uniform
  $\mathbf{A}_t \leftarrow \mathbf{A}_0 + \eta \mathbf{A}_0 \odot \mathbf{N}_{\mathbf{A},t}$ // $\mathbf{A}_t \in [0,1]^{n_1 \times n_1}$
  Sample $\mathbf{N}_{\mathbf{F},t} \sim [-0.5, 0.5]^{n_1 \times k_1}$ // i.i.d, uniform
  $\mathbf{F}_t \leftarrow \mathbf{F}_0 + \eta \mathbf{F}_0 \odot \mathbf{N}_{\mathbf{F},t}$ // $\mathbf{F}_t \in [0,1]^{n_1 \times k_1}$
end
/* generate edges */
$\mathbf{A} \leftarrow \mathbf{0}^{n_1^M \times n_1^M}$ // zero matrix in sparse format
for $i = 1, \ldots,$ num_edges **do**
  for $t = 1, \ldots, M$ **do** $r_t, c_t \leftarrow$ Sample a position in $\mathbf{A}_t$ prop. to its value ;
  $r \leftarrow \sum_{t=1}^{M} r_t \times n_1^{t-1}$ and $c \leftarrow \sum_{t=1}^{M} c_t \times n_1^{t-1}$
  $\mathbf{A}_{rc} \leftarrow 1$ and $\mathbf{A}_{cr} \leftarrow 1$ // add an undirected unweighted edge
end
/* generate attribute memberships */
$\mathbf{F} \leftarrow \mathbf{0}^{n_1^M \times k_1^M}$ // zero matrix in sparse format
for $i = 1, \ldots,$ num_memberships **do**
  for $t = 1, \ldots, M$ **do** $r_t, c_t \leftarrow$ Sample a position in $\mathbf{F}_t$ prop. to its value ;
  $r \leftarrow \sum_{t=1}^{M} r_t \times n_1^{t-1}$ and $c \leftarrow \sum_{t=1}^{M} c_t \times k_1^{t-1}$
  $\mathbf{F}_{rc} \leftarrow 1$
end

Lemma 5 (CKP preserves [P3]). *If $\mathcal{G}_1$ and $\mathcal{G}_2$ obey the spectral radius vs. triangle count power law with exponent γ, i.e., $\sigma_a \propto \Delta_a^{\gamma}$, so does $\mathcal{G}_1 \bar{\otimes} \mathcal{G}_2$.*

The proofs, given in appendix, use the properties of matrix Kronecker product [26] and two key observations: (1) edge count, node count, triangle count and spectral radius of AIS for an attribute a are *explicit* algebraic functions of the adjacency matrix $\mathbf{A}$ and the column in $\mathbf{F}$ which corresponds to a; (2) each column in $\mathbf{F}_1 \otimes \mathbf{F}_2$ is the Kronecker product of a column in $\mathbf{F}_1$ and a column in $\mathbf{F}_2$. Given this, our main result is:

Theorem 6 (SOAR graphs provably obey patterns P1–P3). *If $\mathcal{G}_1 = (\mathbf{A}_1, \mathbf{F}_1)$ obeys patterns P1–P3 with exponents α, β and γ respectively, then $\mathcal{G} = \text{SOAR}(\mathbf{A}_1, \mathbf{F}_1, M, \eta = 0)$ also obeys P1–P3, with the same exponents α, β and γ.*

Proof. We prove this using induction on the number of steps $t = 1, \ldots, M$. It is given that $\mathcal{G}_1$ follows P1–P3, hence the base case for $t = 1$ is true. Now suppose

for $1 \leq t < M$, $\mathcal{G}_t$ follows P1–P3. Then, using Lemmas 3, 4 and 5, $\mathcal{G}_t \bar{\otimes} \mathcal{G}_1 = \mathcal{G}_{t+1}$ follows P1–P3. Thus, by induction, $\mathcal{G} = \mathcal{G}_M$ obeys P1–P3. □

Although Theorem 6 assumes no noise, it can be easily extended to the stochastic version of the SOAR generator to give similar guarantees in expectation. Our simulation studies, presented in Sect. 5.3, confirm our theoretical results.

Discussion. We elaborate on various aspects of the proposed SOAR model. *(a) Input parsimony:* SOAR, belonging to the paradigm of *one-shot graph generation*, has only four knobs to set: two (small) initiator matrices $(\mathbf{A}_1, \mathbf{F}_1)$, number of recursive steps M and noise level η. In contrast, *evolutionary* models typically need knobs for node-arrival, lifetime, sleep-time and linking processes (e.g., [29]). *(b) Attribute correlations:* SOAR implicitly incorporates attribute correlations, as Kronecker product naturally leads to recursive community structure [18]. Contrast this with [24] which explicitly models attribute correlations. *(c) Parameter fitting:* Given a social-affiliation network $\mathcal{G} = (\mathbf{A}, \mathbf{F})$, its parameters for SOAR model can be learned by employing KronFit [18] for $\mathbf{A}$ and $\mathbf{F}$ separately. *(d) Parameter selection:* To create social-affiliation graphs with homophily, we recommend choosing initiators such that the entries of $\mathbf{F}_1\mathbf{F}_1^T$ are correlated with those of $\mathbf{A}_1$. Intuitively, this ensures that nodes with similar attributes are linked in the initiator and the self-similarity of Kronecker product passes this property on to the final graph.

5.3 Simulation Studies

We compare SOAR to two representative baselines – AGM [24] and SAN [14] – which were the most recent works in categories [B] and [C] from Sect. 2. Quantitative experiments compare the time taken by the models to generate graphs of comparable sizes, while qualitative experiments verify whether the models are able to generate graphs obeying the three discovered patterns – **[P1]** Edge count vs. node count power law relation, **[P2]** Triangle count vs. node count power law relation, **[P3]** Spectral radius vs. triangle count power law relation – as well as the following well-known properties: **[P4]** Skewed distributions[2] of #friends per node (node degree), #attributes per node (attribute degree of node) and #nodes per attribute (AIS node count) [29], **[P5]** Skewed distribution of eigenvalues of adjacency matrix [7].

We use the open-sourced code for SAN as is, but adapted AGM to get a skewed distribution of #nodes per attribute (i.e., group size [29]) and subsequently generated edges using the default Fast Chung Lu model. For SOAR, we use initiators from Fig. 6a–b (observe the correlation between $\mathbf{F}_1\mathbf{F}_1^T$ and $\mathbf{A}_1$) replacing $1 \to 0.6, 0 \to 10^{-4}$ for stochasticity (and scaling the remaining entries appropriately), recursive steps $M = 8$ and noise level $\eta = 0.5$. This yields a graph with $0.4M$ nodes, $5.6M$ edges, $65K$ attributes and $2M$ attribute memberships.

[2] Distributions having an asymmetric long or heavy tail, e.g., log-normal, log-logistic.

Quantitative Evaluation. Figure 5 compares generation time of SOAR vs. SAN for five different graph sizes (AGM, due to the explicit enforcing of attribute correlations, scaled poorly with #attributes). Running times are averaged over 10 runs and experiments were performed on Mac OSX Yosemite with 2.7 GHz Intel i5 core and 16 GB main memory. We find that SOAR scales linearly, i.e., slope ≈1 in log-log scale. SAN also shows the desired linear scalability, but was 50× slower for ∼1M edges plus memberships.

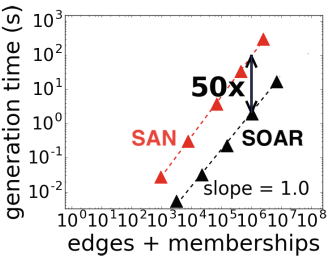
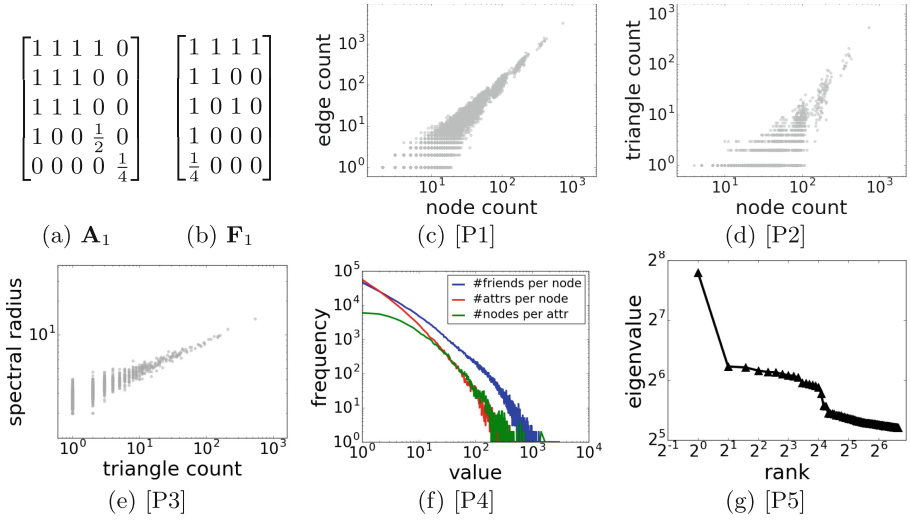


Fig. 5. Speed and scalability.

Qualitative Evaluation. From Figs. 6 and 7, we observe that only the proposed SOAR model is able to generate graphs obeying all these five patterns (Fig. 6), whereas the baselines fail at least one of them (Fig. 7a). In the interest of space, we show only one failed pattern per baseline: AGM leads to very low triangle count for AIS, perhaps due to its undesirably high importance to attribute correlation and homophily, which leads to few edges between nodes sharing attributes when the number of attributes is large (Fig. 7b); SAN produces an almost flat eigenvalue distribution (excluding first three values), likely due to the underlying preferential attachment model (Fig. 7c).



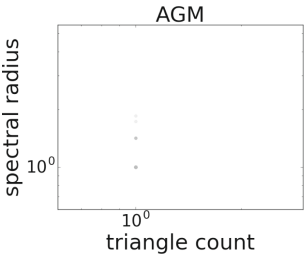
(a) $\mathbf{A}_1$ (b) $\mathbf{F}_1$ (c) [P1] (d) [P2] (e) [P3] (f) [P4] (g) [P5]

Fig. 6. SOAR generates realistic graphs: initiators in (a–b) lead to the discovered patterns P1–P3 (c–e) and skewed degree and eigenvalue distributions P4–P5 (f–g).

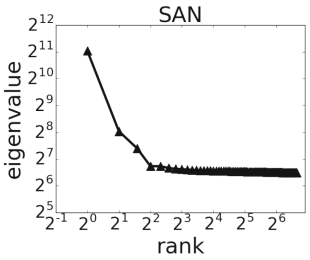
In sum, our simulations demonstrate that SOAR is able to generate social-affiliation graphs obeying all observed patterns in a fast and scalable manner.

Pattern	AGM	SAN	SOAR
[P1]		✓	✔
[P2]		✓	✔
[P3]		✓	✔
[P4]	✔	✔	✔
[P5]	✔		✔

(a)



(b)



(c)

Fig. 7. (a) Graphs generated by baselines (AGM, SAN) disobey at least one pattern, e.g., (b) [P3] of AGM and (c) [P5] of SAN. Here, ✓ denotes empirical adherence based on a few parameters, while ✔ indicates *theoretical* adherence as well. (Color figure online)

6 Conclusion

We investigated the problem of pattern analysis and modeling of social-affiliation graphs – a friendship graph where users have many, binary attributes e.g., check-ins, page likes or group memberships – with the help of four large publicly-available real-world datasets. Our contributions are: (i) **Patterns:** We discovered three new consistent patterns concerning the structural properties of attribute-induced subgraphs and illustrated how the findings can be leveraged for anomaly detection. (ii) **Model:** We proposed SOAR model to produce synthetic social-affiliation graphs *provably* matching all observed patterns. It is based on the principle of self-similarity, implicitly incorporates attribute correlations, scales linearly with graph size and is up to 50× faster than the currently available generators for social-affiliation graphs. Our code is open-sourced at www.github.com/dhivyaeswaran/soar. Similar exploration of node-attributed graphs with categorical/real-valued attributes is a valuable direction for future work.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grants No. CNS-1314632, IIS-1408924 and by DARPA under award FA8750-17-2-0130. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

Appendix (Proofs from Sect. 5)

First, recall the following properties of the Kronecker product [26] for any four suitably sized matrices A, B, C and D: $(A \otimes B)^T = A^T \otimes B^T$; $(A \otimes B)(C \otimes D) = AB \otimes CD$; $\text{Tr}[A \otimes B] = \text{Tr}[A]\text{Tr}[B]$; $\sigma(A \otimes B) = \sigma(A)\sigma(B)$ where $\sigma(\cdot)$ is the spectral radius.

Next, observe that edge count, node count, triangle count and spectral radius of AIS for an attribute a can be explicitly expressed as a function of adjacency matrix $\mathbf{A}$ and the a^{th} column in $\mathbf{F}$ (call it f_a) as follows: (i) Node count of AIS, $n_a = f_a^T f_a$; (ii) Edge count of AIS, $m_a = \frac{1}{2} f_a^T \mathbf{A} f_a$; (iii) Triangle count of AIS, $\Delta_a = \frac{1}{6}\text{Tr}[(\mathcal{D}(f_a)\mathbf{A}\mathcal{D}(f_a))^3]$, assuming no self loops – here, $\mathcal{D}(f_a)$ denotes the diagonalization of vector f_a; (iv) Spectral radius of AIS, $\sigma_a = \sigma(\mathcal{D}(f_a)\mathbf{A}\mathcal{D}(f_a))$.

Let the compact notation, $\bigotimes_{j=1}^{n} A_j$ denote $A_1 \otimes A_2 \ldots \otimes A_n$. Accordingly, every column of $\bigotimes_{j=1}^{n} A_j$ can be expressed as the Kronecker product of a column from each A_j, $j \in \{1, \ldots, n\}$. We are now ready to state our proofs.

Proof (Lemma 3). Any column f_a in $\mathbf{F}_1 \otimes \mathbf{F}_2$ is a Kronecker product of columns $f_{i,1}$ in $\mathbf{F}_1$ and $f_{j,2}$ in $\mathbf{F}_2$ for some i, j. The node count of AIS of a is $f_a^T f_a = (f_{i,1} \otimes f_{j,2})^T (f_{i,1} \otimes f_{j,2})$ which simplifies to $(f_{i,1}^T f_{i,1})(f_{j,2}^T f_{j,2})$ i.e., $n_a = n_{i,1} n_{j,2}$. Similarly, the edge count of AIS of a is $m_a = \frac{1}{2}(f_{i,1} \otimes f_{j,2})^T (\mathbf{A}_1 \otimes \mathbf{A}_2)(f_{i,1} \otimes f_{j,2})$ which can be written as $2(\frac{1}{2} f_{i,1}^T \mathbf{A}_1 f_{i,1})(\frac{1}{2} f_{j,2}^T \mathbf{A}_2 f_{j,2}) \propto m_{i,1} m_{j,2}$. Now, as $\mathcal{G}_1, \mathcal{G}_2$ follow [P1] with exponent α (given), $m_{i,1} \propto n_{i,1}^{\alpha}$ and $m_{j,2} \propto n_{j,2}^{\alpha}$. Hence, $m_a \propto n_a^{\alpha}$.

Proof (Lemma 4). Again, let $f_a = f_{i,1} \otimes f_{j,2}$ for attributes i, j, a in $\mathcal{G}_1, \mathcal{G}_2$ and $\mathcal{G}_1 \otimes \mathcal{G}_2$ respectively. The node count of AIS of a, again, is $n_a \propto n_{i,1} n_{j,2}$. The triangle count of AIS of a is $\Delta_a = \frac{1}{6}\text{Tr}[(\mathcal{D}(f_a)(\mathbf{A}_1 \otimes \mathbf{A}_2)\mathcal{D}(f_a))^3]$ which can be simplified as $\frac{1}{6}\left(\text{Tr}[(\mathcal{D}(f_{i,1})\mathbf{A}_1\mathcal{D}(f_{i,1}))^3]\right)\left(\text{Tr}[(\mathcal{D}(f_{j,2})\mathbf{A}_2\mathcal{D}(f_{j,2}))^3]\right) \propto \Delta_{i,1}\Delta_{j,2}$ using the first, second and third Kronecker properties stated above. Now, as i and j follow [P2] with exponent β (given), $\Delta_{i,1} \propto n_{i,1}^{\beta}$, and $\Delta_{j,2} \propto n_{j,2}^{\beta}$. This results in $\Delta_a \propto n_a^{\beta}$.

Proof (Lemma 5). Once again, let $f_a = f_{i,1} \otimes f_{j,2}$ for attributes i, j, a in $\mathcal{G}_1, \mathcal{G}_2$ and $\mathcal{G}_1 \otimes \mathcal{G}_2$ respectively. We know from the previous proof that triangle count of AIS of a follows $\Delta_a \propto \Delta_{i,1}\Delta_{j,2}$. Now, spectral radius of AIS of a is $\sigma_a = \sigma(\mathcal{D}(f_a)(\mathbf{A}_1 \otimes \mathbf{A}_2)\mathcal{D}(f_a))$ which is $\sigma(\mathcal{D}(f_{i,1})\mathbf{A}_1\mathcal{D}(f_{i,1}))\sigma(\mathcal{D}(f_{j,2})\mathbf{A}_2\mathcal{D}(f_{j,2})) = \sigma_{i,1}\sigma_{j,2}$ due to the second and fourth Kronecker properties stated above. As graphs $\mathcal{G}_1, \mathcal{G}_2$ follow [P3] with exponent γ (given), i.e., $\sigma_{i,1} \propto \Delta_{i,1}^{\gamma}$ and $\sigma_{j,2} \propto \Delta_{j,2}^{\gamma}$. Therefore, $\sigma_a \propto \Delta_a^{\gamma}$.

References

1. Power law. https://en.wikipedia.org/wiki/Power_law
2. Akoglu, L., Faloutsos, C.: RTG: a recursive realistic graph generator using random typing. Data Min. Knowl. Discov. **19**(2), 194–209 (2009)
3. Akoglu, L., Tong, H., Koutra, D.: Graph based anomaly detection and description: a survey. Data Min. Knowl. Discov. **29**(3), 626–688 (2015)
4. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
5. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In: ICWSM. The AAAI Press (2009)
6. Chakrabarti, D., Faloutsos, C.: Graph mining: laws, generators, and algorithms. ACM Comput. Surv. **38**(1), 2 (2006)

7. Chakrabarti, D., Zhan, Y., Faloutsos, C.: R-MAT: a recursive model for graph mining. In: SDM, pp. 442–446. SIAM (2004)
8. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: KDD, pp. 1082–1090. ACM (2011)
9. Crandall, D.J., Cosley, D., Huttenlocher, D.P., Kleinberg, J.M., Suri, S.: Feedback effects between similarity and social influence in online communities. In: KDD, pp. 160–168. ACM (2008)
10. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: SIGCOMM, pp. 251–262 (1999)
11. Fond, T.L., Neville, J.: Randomization tests for distinguishing social influence and homophily effects. In: WWW, pp. 601–610. ACM (2010)
12. Goh, K.I., Kahng, B., Kim, D.: Spectra and eigenvectors of scale-free networks. Phys. Rev. E **64**(5), 051903 (2001)
13. Goldenberg, A., Zheng, A.X., Fienberg, S.E., Airoldi, E.M.: A survey of statistical network models. Found. Trends Mach. Learn. **2**(2), 129–233 (2009)
14. Gong, N.Z., Xu, W., Huang, L., Mittal, P., Stefanov, E., Sekar, V., Song, D.: Evolution of social-attribute networks: measurements, modeling, and implications using Google+. In: IMC, pp. 131–144. ACM (2012)
15. Kim, M., Leskovec, J.: Multiplicative attribute graph model of real-world networks. Internet Math. **8**(1–2), 113–160 (2012)
16. Koutra, D., Koutras, V., Prakash, B.A., Faloutsos, C.: Patterns amongst competing task frequencies: super-linearities, and the Almond-DG model. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds.) PAKDD 2013. LNCS (LNAI), vol. 7818, pp. 201–212. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37453-1_17
17. Leskovec, J., Backstrom, L., Kumar, R., Tomkins, A.: Microscopic evolution of social networks. In: KDD, pp. 462–470. ACM (2008)
18. Leskovec, J., Chakrabarti, D., Kleinberg, J.M., Faloutsos, C., Ghahramani, Z.: Kronecker graphs: an approach to modeling networks. JMLR **11**, 985–1042 (2010)
19. Leskovec, J., Kleinberg, J.M., Faloutsos, C.: Graph evolution: densification and shrinking diameters. TKDD **1**(1), 2 (2007)
20. McGlohon, M., Akoglu, L., Faloutsos, C.: Weighted graphs and disconnected components: patterns and a generator. In: KDD, pp. 524–532. ACM (2008)
21. Mislove, A., Marcon, M., Gummadi, P.K., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: IMC, pp. 29–42. ACM (2007)
22. Newman, M.E.J., Watts, D.J., Strogatz, S.H.: Random graph models of social networks. Proc. Natl. Acad. Sci. **99**(Suppl. 1), 2566–2572 (2002)
23. Perozzi, B., Akoglu, L., Sánchez, P.I., Müller, E.: Focused clustering and outlier detection in large attributed graphs. In: KDD, pp. 1346–1355. ACM (2014)
24. Pfeiffer III, J.J., Moreno, S., La Fond, T., Neville, J., Gallagher, B.: Attributed graph models: modeling network structure with correlated attributes. In: WWW, pp. 831–842. ACM (2014)
25. Rabbany, R., Eswaran, D., Dubrawski, A.W., Faloutsos, C.: Beyond assortativity: proclivity index for attributed networks (ProNe). In: Kim, J., Shim, K., Cao, L., Lee, J.-G., Lin, X., Moon, Y.-S. (eds.) PAKDD 2017. LNCS (LNAI), vol. 10234, pp. 225–237. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57454-7_18
26. Schacke, K.: On the Kronecker product. Master's thesis, University of Waterloo (2004)

27. Titchmarsh, E.C., Heath-Brown, D.R.: The Theory of the Riemann Zeta-Function. Oxford University Press, Oxford (1986)
28. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 440–442 (1998)
29. Zheleva, E., Sharara, H., Getoor, L.: Co-evolution of social and affiliation networks. In: KDD, pp. 1007–1016. ACM (2009)