# A Deductive-Formal Derivation for the Preferential Attachment Metric for Link Prediction

Alane M. de Lima[a], Murilo V. G. da Silva[a], André L. Vignatti[a,b]

[a]*Department of Computer Science, Federal University of Parana, Curitiba, Brazil*
[b]*Corresponding author: vignatti@inf.ufpr.br*

**Abstract**

We present an analytical derivation of the *preferential attachment metric* to predict social ties in complex networks. This metric was originally proposed by Newman in 2001 and by Barabási et al. in 2002 and it was obtained by empirical means. We propose in this paper an analysis based on a *deductive-formal reasoning*, giving the metric a formal theoretical basis. In our analysis we use two random graph models for power-law graphs. We show that in these models, by using formal reasoning, we can derive the preferential attachment metric proposed by Newman and Barabási et al.

*Keywords:* link prediction, preferential attachment, formal analysis.

## 1. Introduction

For a given *complex network* and a pair of its non-adjacent vertices, the link prediction problem asks for the probability that these vertices may become connected by an edge in a near future. Even though this problem has been studied since early 2000's [1], it still remains challenging and has been focus of recent research on the area of complex networks [2, 3]. As pointed by Liben and Kleinberg [1], most of the existing link prediction methods are based on network metrics (e.g. common neighbors, Jaccard measure, Katz centrality). In this work, we are particularly interested in methods based on one of these type of metrics. More precisely, we deal with the *preferential attachment* (PA) metric [1, 4, 5], which is suitable for networks modeled by graphs that have a degree distribution following a power-law function, which are referred here as power-law graphs. Power-law graphs appear in many contexts, including technological, biological, and social networks, so the subject has attracted attention lately from many authors [2, 3, 6]. When first proposed [4, 5], the PA metric was obtained empirically, from experiments on real networks. Indeed, as observed by Liben and Kleinberg [1] in the context of authorship and co-authorship in research publications, where vertices represent researchers and an edge between two vertices $x$ and $y$ indicates that $x$ and $y$ have published a paper together, *"Newman [5] and Barabási et al. [4] have further proposed, on the basis of empirical evidence, that the probability of co-authorship of $x$ and $y$ is correlated with the product of the number of collaborators of $x$ and $y$"*. This raises an issue: from the

point of view of formal sciences (such as logic, mathematics, statistics, theoretical computer science, etc.), the lack of a formal proof is a stumbling block for further results regarding the PA metric.

In this paper we show a deductive-formal derivation of the PA metric. More specifically, we start with three random graph models (as the premises of the formal system), and then we give an analytical derivation for the PA metric in these models. We conclude that the result is considered epistemologically stronger since the result holds both in the formal and empirical paradigms.

## 2. Preliminaries

*Reasoning Methods*: deductive reasoning is the process of reasoning from one or more statements (premises) to reach a logically certain conclusion. If all premises are true and the rules of deductive logic are followed, then the conclusion is necessarily true [7]. Inductive reasoning (as opposed to deductive) is a reasoning in which the premises are seen as providers of strong evidence for the truth of the conclusion [7]. Formal sciences are constructed on formal-deductive arguments (theorems), and inductive reasoning is not accepted. Thus, in the context of a formal epistemology, a PA metric cannot be used unless a proper derivation (i.e., a proof) is provided.

*Models*: A model is a simplified abstract view of a complex reality. Ideally the formal model should not produce theoretical consequences that are contrary to what is found in reality [8, 9]. Typically, however, the model is an approximation, since a complete and precise representation of the reality may be impossible [10]. Thus, it is commonplace that several different models are considered to represent the same reality. In our context, the models are the initial premises of the deductive reasoning process.

*Power-law Graphs*: The empirical study of large real-world networks in the late 1990's and early 2000's [11, 12] showed that the number of nodes of a given degree $i$ is proportional to $i^{-\beta}$, where $\beta > 0$. Such degree distribution is called *power law*. This discovery was soon followed by works describing models for those networks, which involve the use of random graphs with power-law degree distribution, referred here as *power-law graphs*.

*Preferential Attachment (PA) Metric*: This metric was proposed by Newman [5] and Barabási et al. [4] independently, through experiments. Given a pair of vertices $(u, w)$, the PA metric is defined by $\mathrm{PA}(u, w) = \frac{d(u)d(w)}{\sum_v d(v)}$, where $d(v)$ is the degree of vertex $v$ of a given graph.

So, starting from the power-law graph models as the premises, we provide an analysis that has all steps based on deductive reasoning and, in the end, we give a definition of the PA metric.

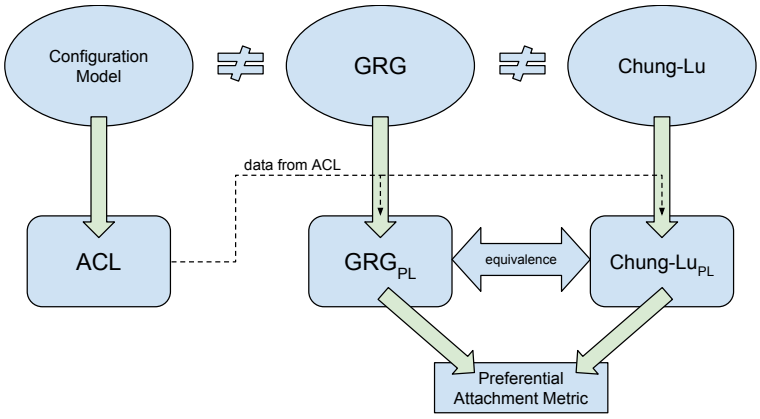


Figure 1: The complete path of the analysis.

## 3. Analysis

Our analysis is an interplay between several models, with theorems supporting the intended result (see Figure 1 for a scheme). As common practice in deductive reasoning, we take the models as premises, from where the theorems are then derived. Next, we briefly describe the models that we used (for an in-depth exposition, see [13]). Throughout this paper, we use $\approx$ to denote an *asymptotic approximation*.

*Configuration Model* [14]: this is a random graph model suitable when we need certain degree distribution of the graph. The model uses an approach where edge connections are probabilistic, but the list of vertex degrees (and therefore the number of edges) is fixed. Let $d(v)$ be the degree of a vertex $v = 1, \ldots, n$. The idea is to create a random graph in the following way. For each vertex $v$, let it have $d(v)$ "half edges" incident to it. Now pick a random perfect matching between the half edges pairing them and connecting them together. As a result we obtain a graph such that each vertex has exactly the desired degree $d(v)$. In many contexts, proving results in this model is not an easy task, since the probabilistic events carry dependencies, and that makes it very hard to obtain useful results, in particular for algorithm design.

*$ACL(\alpha, \beta)$ Model* [15]: the model is obtained from the configuration model using a specific degree distribution for a power-law function. The model depends on input parameters $\alpha$ and $\beta$ where, roughly speaking, $\alpha$ corresponds to the logarithm of the graph size, and $\beta$ is the power-law constant (weakly) related to the graph density. Let $y_i$ be the number of vertices with degree $i$. In this model, by definition, $y_i = \lfloor \frac{e^\alpha}{i^\beta} \rfloor$. Notice that the maximum degree of the graph is $\lfloor e^{\alpha/\beta} \rfloor$ (which is the case where $y_i = 1$). Also, we can ignore rounding [13]. Assuming[1] $\beta > 2$, the number of vertices is $\sum_{i=1}^{e^{\frac{\alpha}{\beta}}} \frac{e^\alpha}{i^\beta} \approx \zeta(\beta) e^\alpha$ and the number of edges is

[1]This is for the purpose of easing the exposure, as [13] deals with other cases for $\beta$.

$\frac{1}{2}\sum_{i=1}^{e^{\frac{\alpha}{\beta}}} i\frac{e^\alpha}{i^\beta} \approx \frac{1}{2}\zeta(\beta-1)e^\alpha$, where $\zeta(t) = \sum_{n=1}^{\infty}\frac{1}{n^t}$ is the *Riemann zeta function.*

*Generalized Random Graph (GRG) Model* [16]: in this model edges are added independently, and weights on the vertices are used to generate arbitrary distribution of the degrees, including the power-law distribution. Starting with a vertex set $V = \{1, 2, ..., |V|\}$, let each $v \in V$ have weight $w_v$. Let $w$ be an array with entries $w_1, ..., w_{|V|}$. In the edge set $E$, every edge $uz$ is created independently with probability $\mathbb{P}(uz \in E) = \frac{w_u w_z}{\ell_n + w_u w_z}$, where $\ell_n = \sum_{v\in V} w_v$. Note that the degree distribution depends on $w$. To obtain a power-law distribution, we build $w$ using the degree distribution from the ACL model, which is a power-law distribution. However, note that while the ACL model is a specification of the *configuration model*, here we are providing a specification of the GRG model. In Section 3.1 we give more details.

*Chung–Lu Model* [17, 18]: this model is similar to the GRG model, the difference relying on the probability $\mathbb{P}(uz \in E) = \frac{w_u w_z}{\ell_n}$, where again $\ell_n = \sum_{v\in V} w_v$. For some values of $w_v$, there is hardly any difference between $\mathbb{P}(uz \in E) = \frac{w_u w_z}{\ell_n + w_u w_z}$ and $\mathbb{P}(uz \in E) = \frac{w_u w_z}{\ell_n}$. In fact, as shown in [13] (Theorem 6.19), when $\sum_{v\in V} w_v^3 = o(n^{3/2})$, GRG and Chung–Lu models are equivalent, but this general condition is not always met for power-law graphs. In Section 3.1, we show how these two models are related.

### *3.1. Deriving the Metric*

Our goal in this section is to derive a formula for the *probability* of an edge appearing between two vertices that is consistent with the empirical definition of the metric.

Due to dependency between events in the configuration model, it is unclear how to obtain a probability for the link prediction directly for the ACL model when it is used as a power-law random graph model. However, since some parameters in the ACL model (e.g., $y_i$, $\alpha$, and $\beta$) are convenient when dealing with power-law graphs, we use them as mathematical tools to specify the parameters for the GRG and the Chung-Lu models. In the GRG model, we set the weights $w$ so that $y_i$ of its entries are equal to $i$, for $i = 1, ..., \Delta$, where $\Delta = \lfloor e^{\alpha/\beta} \rfloor$ (note that, similarly, in the ACL model there are $y_j$ vertices of a given fixed degree $j$). Now, the specific GRG model with power-law parameters creates an edge $uz$ independently with probability $\mathbb{P}(uz \in E) = \frac{w_u w_z}{\ell_n + w_u w_z} \approx \frac{w_u w_z}{\zeta(\beta-1)e^\alpha + w_u w_z}$, where $\ell_n = \sum_{v\in V} w_v$.

**Lemma 3.1** (Vignatti and da Silva (2016) [6], Lemma 2.1)**.** *Let* $w_u, w_v \in \{1, \ldots, e^{\frac{\alpha}{\beta}}\}$. *Then* $e^\alpha\zeta(\beta-1) + w_u w_v \approx e^\alpha\zeta(\beta-1)$.

By Lemma 3.1, $\mathbb{P}(uz \in E) \approx \frac{w_u w_z}{\zeta(\beta-1)e^\alpha} \approx \frac{w_u w_z}{\ell_n}$, which is the probability of creating an edge in the Chung–Lu model. The results then hold on both GRG and Chung–Lu models[2].

**Theorem 3.2.** *Let* $v \in V$ *such that* $w_v = k$. *Then* $\mathbb{E}[d(v)] \approx k$, *where* $d(v)$ *is the degree of* $v$.

[2]The Chung–Lu model can be shown to be asymptotically equivalent to the GRG model under conditions (see [13], Section 6.8.1) that are not necessarily true in our choice of $w$.

*Proof.* Let $W_i$ be the set of vertices with weight $i$, i.e., $W_i = \{v \in V \mid w_v = i\}$ in GRG or Chung–Lu models following the ACL power-law distribution in the assignment of the weights. Let $d_i(v)$ be the number of neighbors of $v$ in $W_i$.

Note that $d_i(v)$ is a binomial random variable with parameters $n = |W_i|$ and $p = \frac{ik}{e^\alpha \zeta(\beta-1)}$. Therefore,

$$\mathbb{E}[d_i(v)] = |W_i| p_{ik} = \left(\frac{e^\alpha}{i^\beta}\right)\left(\frac{ik}{e^\alpha \zeta(\beta-1)}\right) = \frac{k}{i^{\beta-1}\zeta(\beta-1)},$$

which leads to

$$\mathbb{E}[d(v)] = \sum_i \mathbb{E}[d_i(v)] = \sum_i \frac{k}{i^{\beta-1}\zeta(\beta-1)} \approx \frac{k}{\zeta(\beta-1)}\zeta(\beta-1) = k.$$

The first equality is by linearity of expectation and $d(v) = \sum_i d_i(v)$. □

By Theorem 3.2, our model has power-law expected distribution, since it has $\frac{e^\alpha}{i^\beta}$ vertices of expected degree $i$, for each $1 \le i \le \Delta$. Next, we show that in these models the theoretical and empirical definitions of the PA metric are consistent to each other (i.e., it depends on the degrees of the vertices).

**Theorem 3.3** (Preferential Attachment Metric)**.** *Consider $u, z \in V$ in the models GRG or Chung–Lu following the ACL power-law distribution in the assignment of the weights. Then $\mathbb{P}(uz \in E) \approx \frac{d(u)d(z)}{|E|}$, where $d(u)$ and $d(z)$ are the degrees of $u$ and $z$, respectively.*

*Proof.* $\mathbb{P}(uz \in E) = \frac{w_u w_z}{\ell_n + w_u w_z} \approx \frac{w_u w_z}{\ell_n} \approx \frac{d(u)d(z)}{|E|}$, where the first equality is from the definition of the GRG model, the first approximation uses Lemma 3.1, and the second approximation uses Theorem 3.2, noting that $\ell_n \approx |E|$. □

## 4. Conclusion and Future Work

In this work we showed that the PA metric empirically proposed in [5, 4] is the same as the one we derived through our analysis using deductive reasoning. This result gives the metric a formal theoretical basis, so that it can be investigated in the framework of formal sciences. Since many link prediction metrics originate from the inductive paradigm of reasoning, we suggest using deductive analysis on such metrics as future work.

## References


[1] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (2007) 1019–1031.

[2] C. Lee, M. Pham, N. Kim, M. K. Jeong, D. K. Lin, W. A. Chavalitwongse, A novel link prediction approach for scale-free networks, in: Proc. of the 23rd Int. Conf. on World Wide Web, 2014, pp. 1333–1338.

[3] Z. Wang, Y. Wu, Q. Li, F. Jin, W. Xiong, Link prediction based on hyperbolic mapping with community structure for complex networks, Physica A: Stat. Mech. and its Apls. 450 (2016) 609–623.

[4] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, Evolution of the social network of scientific collaborations, Physica A: Statistical Mechanics and its Applications 311 (2002) 590–614.

[5] M. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2001) 025102.

[6] A. Vignatti, M. da Silva, Minimum vertex cover in generalized random graphs with power law degree distribution, Theoretical Computer Science 647 (2016) 101–111.

[7] I. Copi, C. Cohen, D. Flage, Essentials of Logic, Pearson Prentice Hall, 2007.

[8] H. Freudenthal, The Concept and the Role of the Model in Mathematics and Natural and Social Sciences, Springer Netherlands, 1961.

[9] T. Ritchey, Outline for a morphology of modelling methods, Acta Morphologica Generalis 1 (2012).

[10] G. E. P. Box, N. R. Draper, Empirical Model-building and Response Surface, John Wiley & Sons, Inc., 1986.

[11] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the internet topology, SIGCOMM Comput.Com.Rev. 29 (1999) 251–262.

[12] J. Kleinberg, S. Laurence, The structure of the web, Science 294 (5548) (2001) 1849–1850.

[13] R. Hofstad, Random Graphs and Complex Networks, Cambridge, 2016.

[14] E. A. Bender, The asymptotic number of labeled graphs with given degree sequences, Journal of Combinatorial Theory A. 24 (3) (1978) 296–307.

[15] W. Aiello, F. Chung, L. Lu, A random graph model for power law graphs, Experimental Mathematics 10 (2001) 53–66.

[16] T. Britton, M. Deijfen, A. Martin-Lof, Generating simple random graphs with prescribed degree distribution, Journal of Statistical Physics 124 (2006) 1377–1397.

[17] F. Chung, L. Lu, Connected components in random graphs with given expected degree sequences, Annals of Combinatorics 6 (2) (2002) 125–145.

[18] F. Chung, L. Lu, The average distance in a random graph with given expected degrees, PNAS 99 (25) (2002) 15879–15882.