A formal treatment of generalized preferential attachment and its empirical validation

A. Herdağdelen¹, E. Aygün² and H. Bingol¹

¹ Boğaziçi University, Department of Computer Engineering - P.K. 2 TR-34342 Bebek, Istanbul, Turkey
² Istanbul Technical University, Department of Computer Engineering - Ayazağa Kampüsü TR-34469, Istanbul, Turkey

received 3 December 2006; accepted in final form 8 May 2007
published online 4 June 2007

PACS 05.65.+b – Self-organized systems
PACS 89.75.Da – Systems obeying scaling laws
PACS 89.75.Fb – Structures and organization in complex systems

Abstract – Generalized preferential attachment is defined as the tendency of a vertex to acquire new links in the future with respect to a particular vertex property. Understanding which properties influence the link acquisition tendency (LAT) gives us a predictive power to estimate the future growth of network and insight about the actual dynamics governing the complex networks. In this study, we explore the effect of age and degree on LAT by analyzing data collected from a new complex-network growth dataset. We found that LAT and degree of a vertex are linearly correlated in accordance with previous studies. Interestingly, the relation between LAT and age of a vertex is found to be in conflict with the known models of network growth. We identified three different periods in the network’s lifetime where the relation between age and LAT is strongly positive, almost stationary and negative correspondingly.

Introduction. – One of the most profound discoveries in complex-network studies was realizing that the structure and dynamics of many real-world networks do not follow a completely random but rather organized behavior [1]. The power law degree distribution observed in many complex networks has attracted a considerable attention because it is a significant deviation from random behavior [2,3]. In this study, we focus on the dynamics that lead to power law degree distributions.

In a dynamic complex network, there is a continuous creation of vertices and formation of links between the vertices (vertex and link removal can be included in this abstraction as well). For many networks, it is natural to view this process as a competition between the vertices to acquire the newly formed links [4]. The resulting degree distribution will be based on the link acquisition tendency (LAT) values of individual vertices. It is interesting to ask which vertex properties effect link acquisition tendency in which ways.

In this study, we aim to analyze the effect of some basic properties such as age and degree of a vertex on the link acquisition tendency. Such an analysis requires growth data of networks with precise time stamps of the vertex and link creations. As Newman [5] states, obtaining such dynamic data is difficult and most of the time the time resolution is low. Although a limited number of studies analyze the preferential attachment in several networks, their focus is on degree-related preferential attachment [5–9]. Furthermore, many of these reported studies analyze dynamic networks obtained from social collaboration and scientific citation data. Considering these limitations of the previous studies, we decided to use a more generalized methodology that will allow us to analyze not only the effect of degree on link acquisition but also of other vertex properties. Instead of analyzing a previously published dataset, we decided to utilize a new dataset with a high time resolution that will allow us to analyze the preferential attachment in short time scales and comes from a previously unexplored domain.

Methodology. – We assume that the network contains directed links and a vertex is said to acquire a new link if a new link terminating at that vertex is formed. We define the generalized preferential attachment as the tendency to acquire new links with respect to vertex properties [9]. Age and degree of the vertices are the two properties we will discuss. It is possible to formalize the notion of
preferential attachment without referencing a particular vertex property.

Our analysis aims to measure LAT as a function of the vertex properties being investigated. It is based on data collected during an interval $[t_0, t_0 + \Delta t]$ of the network’s lifetime. First, we construct a snapshot graph of the network at $t_0$, record the properties of vertices in that graph, and assume they do not change significantly during the analysis interval $[t_0, t_0 + \Delta t]$. We group the vertices having the same property values together and calculate the average number of new links that each group acquires during the interval. The average number of new links as a function of the vertex property value is a measure of the effect of having a specific property value on the link acquisition tendency. It is possible to view this process as calculating an histogram. We assign each vertex to a bin according to its property value at $t_0$ and record the number of new links accumulated for each bin during the analysis interval. By applying appropriate normalization measures, it is possible to formulate this measure as a probability function conditioned on the property value, as we will see below.

Let $m$ be a generic vertex property (e.g. age, degree, etc.) taking one of the following values $M = \{m_1, m_2, \ldots, m_k\}$ for each vertex. $P(m = m_i)$ is the probability that a vertex has property value $m_i$. This probability distribution is shortly represented as $P(m)$. Let event $L$ denote the acquisition of a new link by a vertex. $P(L)$ is the probability for a particular vertex to acquire a new link. By definition, without any a priori information, $P(L) = 1/n$ where $n$ is the number of vertices. The conditional probability $P(m = m_i|L)$ is the probability of observing a vertex with the property value $m_i$ at the termination point of a newly formed link. This probability distribution function is shortly represented as $P(m|L)$ for notational simplicity. Finally, the conditional probability $P(L|m = m_i)$ is the probability that a particular vertex will acquire the next link to be formed given that the property value of the vertex is $m_i$. It is a measure of the effect of the property $m$ on link acquisition.

By applying the Bayes formula, we can calculate $P(L|m = m_i)$ as follows:

$$ P(L|m = m_i) = \frac{P(m = m_i|L) \cdot P(L)}{P(m = m_i)}. \quad (1) $$

This value gives us the link acquisition tendency as a function of the property $m$ and is a measure of the LAT. Unfortunately, it is not possible to calculate it directly from the data. But we can calculate estimates of measures on the right side of eq. (1) to estimate $P(L|m)$. Let $\Delta t_{total}$ be the total number of links acquired by all vertices during the interval $[t_0, t_0 + \Delta t]$. Let $\Delta m_t$ denote the total number of links acquired by the vertices with property value $m_t$ at $t_0$, during the interval $[t_0, t_0 + \Delta t]$. $P(m|L)$ serves as an estimation for $P(m|L)$. It is calculated as follows:

$$ \hat{P}(m = m_i|L) = \frac{\Delta m_t}{\Delta t_{total}}. \quad (2) $$

If we plug in the empirical estimates of the sample distribution of the property $m$ of the vertices at $t_0$, $\hat{P}(m)$, and $\hat{P}(m|L)$ into the right side of eq. (1), we can obtain an estimate value for the link acquisition tendency as a function of the property $m$.

An important point that is worth being noted is the assumption that the distribution $P(m)$ does not change during the interval $[t_0, t_0 + \Delta t]$. In reality, as time passes, vertex properties such as degree and age change. Previous studies acknowledge this problem and propose using relatively small $\Delta t$ values compared to the lifetime of the network [5–9]. In order to avoid the same problem, we use small $\Delta t$ values and assume that $P(m)$ is stationary during $[t_0, t_0 + \Delta t]$. The effect of different $\Delta t$ values will also be investigated.

A closer examination of our methodology reveals important similarities between the proposed LAT measurement method and the method adopted in [5]. Both methods employ a time window size parameter which regulates the length of the analysis interval and assume that the underlying preferential attachment mechanism is time independent (at least during the analysis interval). One difference is that the final preferential attachment measures reported in this study are relative probability values and it is not possible to compare the results of different analyses (either in time or for different networks) without carrying out a normalization beforehand. The LAT measures reported in this study are normalized conditional probability distributions and hence they have straightforward interpretations. Another difference is that our methodology assumes the degree distribution does not change significantly and uses the distribution sampled at the beginning of the analysis interval $[t_0, t_0 + \Delta t]$ for all calculations regarding that interval while [5] uses the exact distributions for each individual link acquisition incident. Since both methodologies already rely on the assumption that preferential attachment dynamics remain time independent during the analysis, such a simplification in the calculations are quite justifiable and rewarding given the easier normalization techniques.

**Validation.** – Before analyzing the new dataset, we would like to test our new method on a synthetic network built according to the well-known network growth model: the Barabási-Albert (BA) model [2]. Since we know the exact dynamics behind the network growth in the BA model, we can compare our results to the expected ones and see whether our new method correctly captures the dynamics or not.

The BA network is created by setting the model parameters specified in [2] as $t = 1,000,000$, $m_0 = 10$, and $m = 6$. The final network contains approximately one million vertices and six million directed links. The first
Measuring preferential attachment

900000 vertices are used to construct the initial network and the remaining vertices are used to calculate the LAT measure. Our method correctly captures the linear-degree-based preferential attachment as seen in fig. 1(a). It is also analytically known that in the BA model the relation between the birthday of a vertex and the rate at which it increases its degree will be a power law (LAT \( \propto m^{-0.5} \) where \( m \) is the creation time of a vertex) [2]. Our method successfully captures the power law relation as shown in fig. 1(b) along with a dashed line which has the analytically expected slope.

**Data.** – In [10], a new network growth dataset which satisfies the aforementioned requirements is already introduced. The network is constructed by using the data crawled from “Ekiş Sözlük” (literal translation from Turkish is Sour Dictionary) web site [11]. This site, which will be called the Dictionary shortly, is technically a collaborative hypertext dictionary in operation since 15 February 1999. The Dictionary is a site in which one can find explanations and definitions of almost any concept one can think of. Each concept is represented by a title. Each individual definition about a title is called an entry. The entries are listed chronologically under the titles and each entry has an associated timestamp indicating its time of creation. The entries may contain hypertextual cross-references to other (possibly non-existing) titles and they have timestamps indicating the date and time they were written.

By using the raw data, we defined and constructed snapshot graphs of the Dictionary. A snapshot graph, \( G_t \), is a directed graph where each title \( a \) is represented by a vertex \( v_a \) and a cross-reference link from title \( a \) to title \( b \) is represented by an arc from vertex \( v_a \) to vertex \( v_b \). \( G_t \) is constructed by including all vertices and links that were created until \( t \). Several cross-references between the same titles are represented only once by a single arc in the graph. The time resolution of the creation times is one day for the first 2 years and one minute for the rest of the data.

In the end, we obtained a complex-network growth data which contains the vertex and link creation events dating back to the first day of the network (15 February 1999) extending until 01 January 2006. The final network has 1921425 vertices and 6828296 directed links. The degree distribution of the network follows a power law with the following form: \( \hat{P}(m) \propto m^{-\gamma} \), where \( \gamma = 2.126 \) [10].

**LAT as a function of degree.** – The degree-related preferential attachment is an important concept closely related to the degree distributions in networks. The linear preferential attachment hypothesis introduced by the Barabási-Albert (BA) model states that the probability of a vertex to acquire new links is linearly proportional to its current degree [2,6]. Almost all scale-free network models either explicitly incorporate the linear preferential attachment hypothesis [2,4,12,13] or expect it to emerge from the interactions between the growth and dynamics of the network [14,15]. There are some studies, which provide consistent results showing that there is indeed a linear preferential attachment phenomenon in some certain complex networks [5–9,16].

If we let \( m \) represent the degree of a vertex and construct the set \( M = \{m_1, m_2, \ldots, m_q\} \) which is the set of all possible degree values in the network, then the LAT measure we calculate (i.e., \( P(L|m) \)) becomes the degree-related preferential attachment. To measure degree-related LAT, we constructed an initial network by using the snapshot of the network on 01 December 2005 and analyzed the network growth between 01 December 2005 and 31 December 2005. (i.e., \( t_0 \) is 01 December 2005, \( \Delta t \) is 31 days). The calculated LAT values are plotted in fig. 2(a). The findings confirm that link acquisition tendency is linearly proportional to the vertex degree and the best fitting line has the form: \( \text{LAT} = 2.346 \cdot 10^{-6} m + c \).

We carried out the same analysis by using the method proposed by Newman in [5]. Its results are plotted in
Fig. 2: (a) LAT vs. degree calculated by using our newly proposed method and (b) preferential attachment measure obtained by the method adopted in [5], \(t_0 = 01\) December 2005, \(\Delta t = 31\) for both cases.

Fig. 3: Change in the slope of the degree-related LAT value.

December 2005, we repeated measurements by using our original proposed method for different values of \(t_0\) and obtained similar results for every interval we analyzed that is the relation between the LAT and degree of a vertex is linear independent of the time of the analysis. But the slope of the best-fitting line changes significantly as time passes. Figure 3 presents the slope values of the degree-related LAT values which exhibit a significant decrease even for different months in the same calendar year.

LAT as a function of age. – The age of a vertex is defined as the number of days passed since the creation of the vertex. It is possible to consider the creation time (i.e., birthday) of a vertex in the analyses instead of its age and we do so in order to present our graphs in a compatible way with the previous studies [2,17]. In this section, the generic property \(m\) represents the birthday of a vertex in days.

The age-related LAT values calculated from the measurement done between 01 December 2005 and 31 December 2005 are given in fig. 4. The creation times of the vertices are plotted in the \(x\)-axis in days with the first day corresponding to 15 February 1999. The small interval around approximately the 500th day corresponds to a period where the Dictionary was closed temporarily hence no data points exist for that interval.

Interestingly, the age-related LAT does not follow a simple distribution. Instead, we identified three different vertex subsets according to their age values where the link acquisition tendency follows three different distributions. The subsets are named as old vertices, middle vertices, and young vertices. The old vertices are the ones which are created approximately before March 2003 and have birthday values lower than 1000. For the old vertices, the relation between the birthday and link acquisition tendency of a vertex is strongly negative. Therefore,
the earlier a vertex is created the higher LAT value it is expected to have. An exponential model provides a very good fit for LAT values for this period. The best fitting exponential model for the observation has the form $\text{LAT} \propto e^{-3 \cdot 10^{-7} \cdot \text{m}}$.

Young vertices are the ones that are created during the last 60 days prior to the analysis. Among the young vertices, the relation between birthday and link acquisition tendency is positive which means being younger (i.e. having higher birthday values) pays off in terms of LAT value. An exponential model which has the form $\text{LAT} \propto e^{0.015 \cdot \text{m}}$ is the best-fitting exponential model.

The middle vertices are the ones that are created in between the old and young vertices. For the middle vertices, the link acquisition tendency seems to be almost stationary with respect to the birthday. The best-fitting line for the LAT values for this subset has the form $\text{LAT} = -4.583 \cdot 10^{-9} \cdot \text{m} + c$ and it is almost a constant line for all practical purposes.

We should stress that what we are after is not the precise boundaries between these sets of vertices. But the mere recognition of three different subsets of vertices according to their creation times suggests that the relation between the link acquisition tendency and the age of a vertex adopts qualitatively different characteristics during the lifetime of the network.

An analysis of variance (ANOVA) with a significance level of 0.05 confirmed that there is a main group effect of the birthday values: The mean LAT values for the three subsets differ significantly with old vertices being the highest, young vertices the second highest and the middle vertices the lowest.

In order to assess whether this partitioning of the lifetime of the Dictionary is valid not only for a single analysis but whole lifetime of the network, we carried out extensive measurements for different $t_0$ values and all of them yielded similar results. A sample of the age-related LAT values calculated by analyzing different intervals are plotted in fig. 5. Combining the results of ANOVA and the comparative analyses for different intervals, we conclude that partitioning the lifetime of the network into three periods with the aforementioned limits is globally valid for the network. The method adopted in [5] also yields qualitatively similar results which confirm our findings but we are not presenting them here due to space limitations.

In order to analyze the individual characteristics of the three subsets of the vertices in more detail, we carried out 24 different measurements, spanning each calendar month starting from 01 January 2004 to 31 December 2005. For each interval, we calculated the correlation between the age and LAT values of old, middle, and young vertices separately. The average correlation between age and LAT for the old vertices is almost perfectly positive ($r = 0.966$) and has a 95% confidence interval (CI) of $[0.964, 0.967]$. The average correlation between age and LAT for the middle vertices is practically zero: $r = -0.005$ with a 95% CI of $[-0.060, 0.050]$. The average correlation between age and LAT for the late period is strongly negative ($r = -0.528$) and has a 95% CI of $[-0.576, -0.479]$.

The Effect of $\Delta t$. – The only free parameter in the proposed method is the length of the analysis window. We repeated the LAT calculations for degree and age for different values of $\Delta t$ to assess the importance of this parameter. A representative set of our calculations are plotted in fig. 6 and fig. 7. As seen in the figures, the degree-related LAT does not change with respect to differing values of $\Delta t$. For the age-related LAT value, however, the value of $\Delta t$ is more important. For longer $\Delta t$, we cannot observe the increase in the LAT values of the young vertices. This is understandable because for longer time intervals our stationary $P(m)$ distribution

Fig. 4: LAT vs. age, $t_0 = 01$ December 2005, $\Delta t = 31$.

Fig. 5: Age-related LAT for different intervals. (a) 01–31 Jan 2002. (b) 01–31 Dec 2002. (c) 01–31 Jan 2004. (d) 01–31 Dec 2004.
assumption does not hold. The young vertices at the beginning of the analysis are no longer young at the end of the analysis when Δt is one year and this interferes with the calculated LAT values.

Conclusions and future work. – As in compliance with previous studies, the relation between the LAT value and degree of a vertex is found to be linear. The more connected vertices are more likely to acquire new links in the future and this likelihood is a linear function of vertex degree.

The relation between the LAT value and age of a vertex is more complicated. For the old vertices that are created roughly before the 1500th day (which corresponds to some time around March 2003) the LAT value is positively correlated with the age. For the vertices that are created after March 2003 (i.e. middle vertices), the relation between the LAT and age (or creation time) disappears except for the vertices that are created during last 60 days prior to the analysis (i.e. young vertices). For the young vertices being younger pays off in terms of LAT.

In the real-life dynamics, the age of a vertex certainly does not play an explicit role in link acquisition. It is obvious that the users are not inclined towards giving cross-references to other titles just because they are created early. This reasoning suggests that at least one mediator variable should be present and affecting LAT through age. The nature of this (or possibly these) variable(s) calls for future research.

***

This work was partially supported by Bogazici University Research Fund under the grant number 06A105. The authors thank Muhittin Mungan for stimulating discussions on the manuscript.

REFERENCES

[1] Albert R. and Barabasi A. L., Rev. Mod. Phys., 74 (2002) 47.
[2] Barabasi A. L. and Albert R., Science, 286 (1999) 509.
[3] Newman M. E. J., SIAM Rev., 45 (2003) 167.
[4] Pennock David M., Flake Gary W., Lawrence Steve, Glover Eric J. and Giles C. Lee, Proc. Natl. Acad. Sci. U.S.A., 99 (2002) 5207.
[5] Newman M. E. J., Phys. Rev. E, 64 (2001) 025102(R).
[6] Barabasi A.-L., Jeong H., Neda Z., Ravasz E., Schubert A. and Vicsek T., Physica A, 311 (2002) 590.
[7] Jeong H., Neda Z. and Barabasi A. L., Europhys. Lett., 61 (2003) 567.
[8] Redner S., Phys. Today, 58 (2005) 49.
[9] Camille Roth, Generalized preferential attachment: Towards realistic socio-semantic network models, in ISWC 4th International Semantic Web Conference, Workshop on Semantic Network Analysis, Galway, Ireland, November 2005, CEUR-WS Series, Vol. 171, ISSN 1613-0073, pp. 29-42.
[10] Herdağdelen A., Aygün E. and Bingöl H., Measuring generalized preferential attachment in an online hyper-textual dictionary: Eksi Sozluk, in ECCS European Conference on Complex Systems, Paris, France, November 2005.
[11] Eksi Sozluk, http://sozluk.soutimes.org.
[12] Kravivsky P. L., Redner S. and Leyvraz F., Phys. Rev. Lett., 85 (2000) 4629.
[13] Dorogotsev S. N., Mendes J. F. F. and Samukhin A. N., Phys. Rev. Lett., 85 (2000) 4633.
[14] Klemm Konstantin and Eguiluz Victor M., Phys. Rev. E, 65 (2002) 036123.
[15] Vazquez Alexei, Phys. Rev. E, 67 (2003) 056104.
[16] Eisenberg E. and Levanon E. Y., Phys. Rev. Lett., 91 (2003) 138701.
[17] Dorogotsev S. N. and Mendes J. F. F., Phys. Rev. Lett., 62 (2000) 1842.