

Measuring Preferential Attachment in a Hyper-Textual Dictionary Reference Network: Eksi Sözlük^{*}

Amac Herdagdelen¹, Eser Aygun², Haluk Bingol¹

¹ Department of Computer Engineering, Bogazici University

Phone: +90 212 359 45 23-24, Fax: +90 212 287 24 61

Istanbul, TR-34342, Turkey

yalanciborsaci@yahoo.com, bingol@boun.edu.tr

² Department of Computer Engineering, Istanbul Technical University

Istanbul, TR-34469, Turkey

aygunes@itu.edu.tr

Abstract. Existing methods used to analyze the relation between preferential attachment and node degree make use of time dependent measures, which result in limited ability to analyze the temporal characteristics of networks. We introduce time independent measures, which allow us to analyze the networks' preferential attachment behavior in a more precise manner. The two different methodologies are compared on a new complex network data: Eksi Sözlük, which spans the whole lifetime (six years) of a complex network with very precise recordings of the node and edge addition events (i.e. one minute). The relation between the likeliness to receive new links and the present degree of a node is found to be linear. Analyses suggest that time independent measures are better in capturing the dynamics of the network and in some cases, provide results that are very hard to obtain by existing methodologies.

Keywords. Preferential attachment, network growth models, temporal analysis, network dynamics, link acquisition.

^{*} This work was partially supported by Bogazici University Research Projects under the grant number 05A105.

1 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]). Most of the complex networks seem to share some basic properties such as power-law degree distribution, small average path lengths, and high clustering. Among them, the power-law degree distribution has attracted a considerable attention because it is a significant deviation from random behavior and it serves as a basis for the scale-free network concept ([1], [2]). In this study, we focus on the dynamics that lead to power-law degree distributions.

The ubiquity of power-law degree distribution led the scientists to the research of finding out why networks from so diverse origins show the same characteristics and understanding its implications. A line of this research is building up analytical models and comparing them with the real-world data. The original Barabási-Albert (BA) model ([1]) and its generalizations (see [3] for a discussion), growing network model ([4]), growth and deactivation model ([5]), and local models proposed by [6] are among these studies. It is worth noting that in these modeling efforts, an important aim is to be able to model the dynamics of the network, hence they are called growth models and they do not only try to capture the structural properties but also the dynamic evolution of the networks in time.

An important concept introduced by the BA model is the preferential attachment hypothesis and it has deep connections with the power-law distribution. According to the hypothesis, a node with a relatively high degree will be more likely to acquire new links in the future. Almost all scale-free models either assume the existence of preferential attachment and explicitly incorporate it ([1], [3], [4]) or expect it to emerge from the interaction between the growth and dynamics of the network ([5], [6]). This common adoption is of no coincidence because as [4] and [7] analytically show that for the BA model and its generalizations, the nature of the resulting degree distribution depends on the nature of preferential attachment.

Our motivation for conducting this study was observing the fact that although a number of studies form the theoretical relation between the nature of preferential attachment and power-law degree distribution (i.e. being scale-free), the empirical validations that confirm such a tendency is really present in real-world complex networks are relatively low in number. However, there are some studies which provide consistent results showing that there is indeed a preferential attachment phenomenon in some certain complex networks ([8], [9], [10], [11], and [12]). Carrying out such a study is not trivial because for a through analysis of the preferential attachment, the network growth data should contain the exact timestamps of link and node addition events. Any problem with the “quality” of the data in that sense will cause degradation in the reliability of the results. In [8], for instance, the time resolution of the data is one year, which means that the relative ordering of the link additions during each year is unknown. In [11], on the other hand, the full relative ordering of the link additions is known but the actual dates are missing and this fact results in the assumption that all of

the link additions are uniformly distributed in the time line. A detailed discussion for these two and the other studies is made in Section 2.2.

In this study, we tried to measure and see if there is a preferential attachment tendency in a complex network which comes from a slightly different domain than the networks analyzed in the previous studies and has some features missing in other ones (i.e. more precise time resolution). While doing so, we adopted new methodological approaches to provide measures that are more reliable. The main contribution of the new methods for preferential attachment measurement is that they can be classified as time independent measures in contrast to the presently used time dependent measures. Time dependency means that the measures are specific to a certain time interval along the network's lifetime and it is impossible to integrate two different measures obtained from two different time intervals. In contrast, time independent measures provide values that are independent of the specific time interval on which they were calculated and this enables us to integrate different measures obtained in different times. The most obvious application for such measures is carrying out the analyses on very short consecutive time intervals and then integrating the results altogether.

2 Theory and Previous Studies

2.1 Theory

Directed Graph (G): A directed graph G is defined as $G = (V, E)$, where V and E are the set of vertices (nodes) and the set of edges (links) of G , respectively. For two vertices $v_{source}, v_{target} \in V$, an edge $e \in E$ is defined as $e = (v_{source}, v_{target})$.

In-Degree, Out-Degree, Total Degree ($k_i^\downarrow, k_i^\uparrow, k_i^\updownarrow$): For a vertex v_i , out-degree is the number of all edges leaving v_i , in-degree is the number of all edges pointing to v_i , and total degree is the sum of out-degree and in-degree of v_i . These are represented by k_i^\uparrow , k_i^\downarrow , and k_i^\updownarrow , respectively. In this paper, in-degree is called as degree, and represented by k_i , for simplicity.

Degree Distribution ($P(k)$): Degree distribution of a graph shows the probability of having a certain degree k for a randomly selected node of the graph.

The power-law degree distribution, which is a characteristic property of scale-free networks, is defined as $P(k) = ck^{-\lambda}$ where c is a normalization coefficient to ensure $\sum_k P(k) = 1$ and $\lambda > 0$ is called scaling factor of the distribution. In **Figure 1**, the power-law distribution is compared with a typical exponential ($P(k) = ce^{-mk}$) distribution.

The following temporal definitions are used to represent an evolving network in a formal way:

Graph History (Γ): A graph history Γ is defined as $\Gamma = (\Xi, E, \tau^\Xi, \tau^E)$, where Ξ and E are the set of all vertices and the set of all edges created in Γ , respectively, τ_i^Ξ is the creation time of vertex ξ_i , and τ_i^E is the creation time of edge ε_i .

Recent Degree Increase ($rdi_{t,i}$): For a vertex ξ_i and duration Δt , recent degree increase at time t is the number of all edges pointing to ξ_i that are created in the time interval $[t - \Delta t, t)$, and it is represented by $rdi_{t,i}$.

Snapshot Graph (G_t): A snapshot graph G_t of a graph system $\Gamma = (\Xi, E, \tau^\Xi, \tau^E)$ at time t is defined as $G_t = (V_t, E_t)$, where $V_t = \{\xi_i \in \Xi : \tau_i^\Xi < t\}$ and $E_t = \{\varepsilon_i \in E : \tau_i^E < t\}$. For a snapshot graph G_t , the in-degree, out-degree and total degree of a vertex $\xi_i \in V_t$ is represented by $k_{t,i}^\downarrow$, $k_{t,i}^\uparrow$, and $k_{t,i}^\updownarrow$ respectively. Similarly, in-degree is called as degree, and represented by $k_{t,i}$, for simplicity.

Age ($a_{t,i}$): In a graph system, the age of a vertex ξ_i at time t is defined as $a_{t,i} = t - \tau_i^\Xi$.

Preferential Attachment Measures ($l_{t,i}$, $l_t^{degree}(k')$, and $l_t^{age}(a')$): $l_{t,i}$ is the number of new links that vertex ξ_i is expected to gain in the time interval $[t, t + \Delta t)$. By using this measure, it is possible to construct preferential attachment tendency as functions of degree and age. $l_t^{degree}(k')$ gives the expected number of new links to be acquired for a random node with a specified degree. It is defined as:

$$l_t^{degree}(k') = \frac{\sum_{x \in X} l_{t,x}}{|X|} \quad (X = \{x \mid k_{t,x} = k'\}) \quad (1)$$

Similarly $l_t^{age}(a')$ is defined as:

$$l_t^{age}(a') = \frac{\sum_{y \in Y} l_{t,y}}{|Y|} \quad (Y = \{y \mid a_{t,y} = a'\}) \quad (2)$$

Although we cannot directly determine these tendencies analytically for a given data, it is possible to come up with empirical computations, which may serve as approximate measures. The empirically computed values are represented by $\hat{l}_{t,i}$, $\hat{l}_t^{degree}(k')$, and $\hat{l}_t^{age}(a')$, correspondingly, and details of their computation is given in Section 3.2.

2.2 Relevant Studies

It is possible to present the relevant studies under two headings. First group is the modeling efforts for complex networks, and the second is the search for empirical evidence supporting or falsifying these models.

Models: In this study, we focused on two network growth models. The first model is the original BA model ([1]). According to BA model, the network starts from a small set of fully connected core nodes, and new links are formed only between new arriving nodes and the existing nodes. Whenever a new node is introduced to the graph, it forms a link with vertex v_i according to the following probability:

$$p(k_i) = \frac{k_i^\gamma}{\sum_j k_j^\gamma} \quad (3)$$

where the parameter γ is introduced to account for the generalizations of BA model ([3]), and in the original case $\gamma = 1$.

The original BA model assumes a linear dependency between the likeliness to acquire a new connection and the present number of connections. This linearity is of paramount importance as it is shown in [4] and [7] analytically and in [1] empirically that in the presence of sub-linear dependency (i.e. $\gamma < 1$) the degree distribution becomes a stretched exponential; and in the case of super-linear dependency (i.e. $\gamma > 1$), a “winner takes it all” situation is observed and one node acquires all new coming links resulting in a star-like topology. A prediction of BA model is a positive correlation between the age of a node ($a_{t,i}$) and its degree ([5]). The reason is intuitively obvious that the older a node is, the more time it had to acquire links, which results in a higher probability of getting new links. Another important point about the BA model is that it generates an undirected graph (i.e. all edges are symmetric).

The second model is the growth and deactivation model described in [5]. The growth of the network is described by directed links (unlike the BA model). An important concept introduced by this model is the active/inactive node difference. At a given time, there exist a constant number m of active nodes, and an active node always receives an incoming link from a newly arriving node. An inactive node, on the other hand, cannot receive any links. Whenever a new node is introduced to the system, it forms new links pointing to the current active nodes, it is tagged as an active node, and one of the active nodes is deactivated randomly so the number of active nodes is kept constant. The deactivation probability for each active node v_i is inversely proportional to its current number of incoming links:

$$P_i^{\text{deactivation}} \propto (k_i + a)^{-1} \quad (4)$$

where a is a constant bias. A strong simplification of the model is that once a node is deactivated there is no way for it to be re-activated and receive new links. There are a number of predictions associated with this model. First, it is analytically shown that the average increase in the degree of a node is linearly proportional to the present degree ($l_{t,i} \propto k_{t,i}$), which leads to the linear dependence $l_i^{\text{degree}}(k') \propto k'$. This dependence is not explicitly incorporated into the model but is an emergent property of degree-dependent deactivation dynamics ([5]). Secondly, in sharp contrast to the BA model, this model predicts a negative correlation between the age and probability of acquiring a new link. No matter how high degree a node has, as time passes the probability of being deactivated at some point increases. The dependence between the age and probability

of getting a new connection is shown to be compliant with power-law. We defined the term recent degree increase ($rdi_{i,t}$) in correspondence with this study. While the original study does not contain such a concept, we believe the recent degree increase of a node is in close relation with its status of being active or inactive. Active nodes tend to have high recent degree increases; inactive nodes tend to have low (actually zero).

Empirical Studies: In [8], preferential attachment in two different networks is investigated. The networks are the co-authorship networks in fields of mathematics and neuroscience between years 1991 and 1998. The authors use a measure very similar to $l_t^{degree}(k')$ but the normalization according to the number of nodes that have degree k' is not used and the measurements are carried on a fixed time interval of one year. The findings suggest that the relation between degree of a node and its expected rate of link acquisition is linear. However, as Barabási et al. indicated themselves, the data is not complete but only a recent portion of the all co-authorship network is available. Another problem is the time resolution of the network data. Only the years of collaborations are known so the relative ordering of the link formations during each year is missing. Methodologically, the time period of analysis (Δt) has to be kept relatively short because as the nodes continue to acquire new links, their actual degrees increase while the records used for analyzing the relation is fixed to the analysis of time t . Using longer periods may introduce unwanted bias in this sense, and using shorter periods create fluctuations in the results that are hard to eliminate.

The second study relevant to our work is [11], which also focuses on two networks of co-authorships, one in physics the other in biology and medicine. The data used in this study contains the entire relative ordering of node and link additions but the exact timestamps are missing. Therefore, it is assumed that the links are introduced uniformly in the timeline. This study confirms that the preferential attachment tendency is linearly dependent on the current number of links of the nodes. The methodology is very similar to the one in [8] but the normalization of the expected number of new links according to the number of nodes is employed so the measure calculated is virtually the same with $l_t^{degree}(k')$. This normalization is important because for small values of k' , there may be many nodes with k' links, but for high values of k' there are only a few of them if not zero at all. Just summing these values without considering how many nodes contribute to it may introduce some bias in the calculation of tendency to acquire new links, favoring the lowly connected nodes because their number is supposed to be higher than the highly connected ones. However, this kind of normalization does not help to eliminate the problem associated with the length of the time period analyzed.

[10] follows the same methodology, which is adopted in [8] but extends the analyses to different networks from different origins, namely, science citation network, Internet, actor collaboration and scientific co-authorship. The time resolution of the analysis is one year for all of the networks. For the first two a linear dependence for preferential attachment is observed while for the latter two, the dependence is sub-linear.

Of the other two studies, [9] is of significance because it extends the empirical findings to a protein-protein interaction network that has a very diverse origin of the other networks considered so far. [12], on the other hand, employs similar techniques to [8] and [10] but uses a more complete dataset: The full citation network of Physical Review journals, spanning a period from 1893 until 2003. The time resolution of the

analysis is again one year. Both studies confirm a linear dependency between likeliness to acquire a new link and the current number of links.

3 Data and Methodology

3.1 Data

The network that is analyzed in this study is constructed by using the data crawled from “Ekşi Sözlük” (literal translation from Turkish is Sour Dictionary) web site. Ekşi Sözlük (which will be called Sözlük, shortly) is technically a collaborative hypertext dictionary in operation since 15 February 1999, and it is gained a wide popularity among the Internet users in Turkey ([13]).

Sözlük is a dictionary in which one can find explanations and definitions of almost any concept one can think of. In the Sözlük’s jargon, a concept for which information can be found is called a “title” (literal translation of “başlık” from Turkish). Each individual definition, explanation, or information of any kind is called an “entry”. There may be several entries placed under a title. What makes Sözlük different from any other plain text based dictionary is that it contains hyper-textual references to other titles. For instance, in an entry below the title “Stargazer”, which is the name of a masterpiece song of a hard rock band ([14]), the name of the band “Rainbow” can be a hypertext. When clicked on, it may lead to the title “Rainbow”, where entries about both the band and the literal definition of a rainbow can be found.

Sözlük is also a collaborative work of its registered users, who are called “susers” (stands for Sözlük USER), and are eligible to add new titles or enter new entries under existing titles.

The graph history $\Gamma_{\text{Sözlük}}$ is constructed by using the data crawled from Ekşi Sözlük web site. The crawling process consisted of fetching each individual entry, parsing it for possible cross-references, and recording their timestamps. This operation took 168 computing hours and the resulting data is given in **Table 1**.

The graph history corresponding to Sözlük, $\Gamma_{\text{Sözlük}} = (\Xi, E, \tau^{\Xi}, \tau^E)$, is constructed as follows: There is a one-to-one correspondence between the vertices of $\Gamma_{\text{Sözlük}}$ (i.e. elements of Ξ) and the titles of Sözlük. For an edge $\varepsilon_i = (\xi_{\text{source}}, \xi_{\text{target}})$, $\varepsilon_i \in E$ if and only if there is at least one cross-reference from the title *source* to the title *target*. τ^{Ξ} and τ^E are constructed such that τ_i^{Ξ} of ξ_i is the timestamp of the first entry of title *i*, or the timestamp of the first cross-reference pointing to title *i*, whichever is earlier; τ_i^E of edge $\varepsilon_i = (\xi_{\text{source}}, \xi_{\text{target}})$ is the timestamp of the first entry that contains a cross-reference from title *source* to title *target*. From now on, the snapshot graphs of $\Gamma_{\text{Sözlük}}$ will be referred as G_t , for simplicity.

Unlike most of the other complex network examples, Sözlük data spans the whole lifetime of the network since the first day and first node. In addition, the time resolution of the data is very precise: One day for the first two years, and one minute for

subsequent years. Another important characteristic of Sözlük network is that it is not only growing but it is also evolving because of the moderation activities, yet we were able to crawl only the visible entries but not the removed ones.

3.2 Methodology

Our analysis of the data is composed of three parts. In the first part, we employ the methodology used in [11], and for time t we empirically calculate the preferential attachment measures of all nodes ($\hat{l}_{t,i}$) by calculating the current degrees ($k_{t,i}$), and the number of new links they acquired during time interval $[t, t + \Delta t)$ ($\Delta k_{t,i}$).

By the definition, $\Delta k_{t,i}$ is an empirical calculation of the preferential attachment measure $l_{t,i}$, thus $\hat{l}_{t,i} = \Delta k_{t,i}$. Accordingly, preferential attachment tendency as a function of degree is calculated as

$$\hat{l}_t^{\text{degree}}(k') = \frac{\sum_{x \in X} \Delta k_{t,x}}{|X|} \quad (X = \{x \mid k_{t,x} = k'\}) \quad (5)$$

and preferential attachment tendency as a function of age is calculated as

$$\hat{l}_t^{\text{age}}(a') = \frac{\sum_{y \in Y} \Delta k_{t,y}}{|Y|} \quad (Y = \{y \mid a_{t,y} = a'\}) \quad (6)$$

To keep the bias as low as possible during these analysis, it is necessary to use a relatively short time period (i.e. Δt), which is one day for our data, and one year for [8] and [11]. This part of analysis provides us the time dependent measures.

In the second part, we introduce another measurement method that makes use of the whole temporal data, and provide a time independent measure for the preferential attachment tendency. This way, it is possible to tackle with the bias problem faced in [8] and [11] by using very short time periods (e.g. one week) to calculate the number of new links acquired and then summing them altogether for a longer period (e.g. one year). The major problem for coming up with a time independent measure is that, while having a certain number of links, say 100, for year 2001 may provide a relative advantage to acquire new links, in year 2005, having 100 links may not mean that much because all of the nodes are continuously acquiring new links. Previous studies limited themselves with relatively short periods (i.e. one year) to count the new number of new links, because it is assumed that the bias introduced during one year is negligible. What we have to do is to come up with new measures which aim to reflect the same tendency as $l_t^{\text{degree}}(k')$ and $l_t^{\text{age}}(k')$ and is independent of the absolute value of k' but its relative magnitude. This way, it will be possible to integrate the different preferential attachment measures of different snapshot graphs of Sözlük instead of focusing on only one.

The method we adopt is to normalize the measures linearly for each snapshot graph G_t , such that they fall in interval $[0,1]$. The normalization process for a measure χ is as follows:

$$\chi_{t,i}^{norm} = \frac{\chi_{t,i} - \chi_t^{\min}}{\chi_t^{\max} - \chi_t^{\min}} \quad (7)$$

where χ is can be substituted with k , rdi , a , and Δk . The corresponding symbols k_t^{\min} (k_t^{\max}), rdi_t^{\min} (rdi_t^{\max}), a_t^{\min} (a_t^{\max}), and Δk_t^{\min} (Δk_t^{\max}) are the smallest (highest) degree, recent degree increase, age, and the number of new links values of the nodes of graph, respectively. Substituting the new normalized values in Equation (5) and (6), we obtain the new functional representations of preferential attachment tendency:

$$\hat{l}_t^{\deg ree}(k') = \frac{\sum_{x \in X} (\Delta k_{t,x})^{norm}}{|X|} \quad (X = \{x \mid k_{t,x}^{norm} = k'\}) \quad (8)$$

$$\hat{l}_t^{age}(a') = \frac{\sum_{y \in Y} (\Delta k_{t,y})^{norm}}{|Y|} \quad (Y = \{y \mid a_{t,y}^{norm} = a'\}) \quad (9)$$

The time-independent measure of preferential attachment tendency as a function of degree and age is obtained by averaging the time-independent measures calculated for successive snapshot graphs of $\Gamma_{Sözlük}$:

$$\eta^{\deg ree}(k') = \frac{\sum_{t \in T} \hat{l}_t^{\deg ree}(k')}{|T|} \quad (T = \{t_0, t_0 + \Delta t, t_0 + 2\Delta t, \dots, t_{end}\}) \quad (10)$$

$$\eta^{age}(a') = \frac{\sum_{t \in T} \hat{l}_t^{age}(a')}{|T|} \quad (T = \{t_0, t_0 + \Delta t, t_0 + 2\Delta t, \dots, t_{end}\}) \quad (11)$$

where $t_0 > 0$ is a given time for starting the analysis and t_{end} is the timestamp of the last event represented in $\Gamma_{Sözlük}$.

The third part of our analysis aims to present an example application for the new time-independent measures. The results we obtain are used to compare and evaluate two network growth models (BA and growth and deactivation) to see which one provides a better explanation for the data. The methodological tools used are the correlation and partial correlation analysis. By correlation analysis, we calculate the pairwise Pearson correlation coefficients between $k_{t,i}^{norm}$, $\hat{l}_{t,i}^{norm}$, $rdi_{t,i}^{norm}$, $a_{t,i}^{norm}$ obtained for each node for each time step t ; and carry out a significance test to see whether the observed correlations are indeed significant. By partial correlation analyses, we measure the correlation between normalized age $a_{t,i}^{norm}$, and the $\hat{l}_{t,i}^{norm}$ when the effect of current number of links $k_{t,i}^{norm}$ is removed. In this case, BA model predicts a zero correlation (given the degrees of two nodes are equal, the age of a node does not have an effect on the link acquisition), but growth and deactivation model predicts a negative correlation as discussed in Section 2.2. Also the partial correlation between $k_{t,i}^{norm}$ and $\hat{l}_{t,i}^{norm}$ is

calculated when the effect of the recent degree increase ($rdi_{t,i}^{norm}$) is removed. BA model predicts a positive partial correlation because according to the dynamics, even if recent degree increase values of two nodes are the same, the more connected one has greater chances to acquire new links. On the contrary, growth and deactivation model predicts a negative correlation because the nodes with low recent degrees are the inactive nodes, and they cannot receive new links anymore. Similarly, the partial correlation between $rdi_{t,i}^{norm}$ and $\hat{l}_{t,i}^{norm}$ is calculated when the effect of the node degree ($k_{t,i}^{norm}$) is removed. BA model does not predict a strong positive correlation, whereas growth and deactivation model predicts a strong positive correlation because the active nodes, which are more likely to receive new links, tend to have high recent degree increase values independent of their overall degree.

4 Results

Degree Distribution: The degree distribution of the snapshot graph $G_{t_{end}}$ is given in **Figure 2**. We tried also to fit a power-law model, and the scaling factor of the model is calculated as 2.63.

4.1 Preferential Attachment, Degree and Age Dependency

Unnormalized measures: For the calculation of $\hat{l}_t^{degree}(k')$ and $\hat{l}_t^{age}(a')$ values, Δt is set as one year, and t_0 is chosen as equal to $t_{end} - \Delta t$. Degree and age values were normalized in interval $[0,1]$ with a resolution of 0.01. The plot of the values is given in **Figure 3**. The robust best fitting lines for degree and age measures are also superimposed on the corresponding plots. In this configuration, the root mean square error of the robust best line fitting for degree data $RMSE_{degree}$ is 0.060 and for age data $RMSE_{age}$ is 0.227.

Normalized measures: For calculating the normalized preferential attachment measures $\eta^{degree}(k')$ and $\eta^{age}(a')$, Δt was set as one week, and the starting time of analysis, t_0 , was set as equal to $t_{end} - T$, where T is four years. Thus, a snapshot graph for each week between the years 2003 and 2004 was constructed to calculate corresponding time independent measures $\hat{l}_{t,i}^{norm}$ and $a_{t,i}^{norm}$, which in turn provided the time independent measures $\eta^{degree}(k')$ and $\eta^{age}(a')$. The plot of these values with respect to the normalized degree and age values are given in **Figure 4**. The robust best fitting lines for those values are also superimposed on the corresponding figures. Note that, fitting of the age data excludes values before 0.2 because of the “today effect”, which will be discussed later. In this second configuration, the root mean square errors, $RMSE_{degree}^{norm}$ is 0.050 and $RMSE_{age}^{norm}$ is 0.192.

4.2 Interdependency Between Age, Degree and Recent-Degree

In Section 4.1, it is explained how the normalized measures are obtained. Shortly, for each interval of one week $[t + \Delta t)$, $k_{t,i}^{norm}$, $\hat{l}_{t,i}^{norm}$, $a_{t,i}^{norm}$, and $rd_{t,i}^{norm}$ values are calculated for each node in the corresponding graph G_t . Grouping this quadruple for each individual node, a list was formed and a random sub-sampling was carried on this list. As a result, a subset of 2370 elements was selected. The pairwise Pearson correlation coefficients of the measures are given in **Table 2**.

Partial correlation coefficients as defined in Section 3.2, are given in Table 3.

5 Discussion

In accordance with the previous results, we were able to confirm a linear dependency between the preferential attachment tendency and the degree of the nodes. A linear model provides a better robust fit for the normalized measures than the unnormalized ones. ($RMSE_{degree}^{norm} = 0.050 < RMSE_{degree} = 0.060$). Assuming the actual relation is indeed linear, this fact suggests that the normalized method is better in capturing the linearity. However, as it can be seen visually in **Figure 4**, for larger degree values, the preferential attachment relation follows a non-linear form. The same effect is not observed for the unnormalized measures. It is an open question whether this observation is a side effect of the normalization or the preferential attachment really follows a non-linear relationship for large degree values. We were not able to answer it with the current findings.

Another interesting result was obtained by measuring the relation between the age and preferential attachment of a node. With the normalized measures, we observed that the relation between age and number of new links could be modeled by an exponential model ($l_{t,i} \propto e^{9.751a_{t,i}}$). That is, the likeliness to acquire new links for a node increases, as the node gets older. Interestingly, for smaller age values the relation adopts a power law (scaling factor is 0.448) and follows an inverse relation: In the interval $[0,0.2]$, the younger a node the more probable that it will receive a new link in the future. Our account for this phenomenon is the effect of a facility called “Today’s Titles” that lists the titles, which have received new entries during that day. Apart from manual search and following hyperlinks, this facility is the only way to see a group of titles and acts as a buffer in the sense once a new title is created it is immediately put in this list and becomes visible to the other users for a brief time. As time passes the probability that it will be removed from the list (because of new comers) increases. We call this effect as “today effect” and it serves as a good example for the benefits of using more precise measurement methods. While it is possible to visually observe the same effect for unnormalized measures, we were unable to fit a meaningful model (i.e. neither exponential nor power-law) to the data in interval $[0,0.2]$, which suggests that the unnormalized measures are not capable of reflecting dynamic preferences in order of weeks and days.

The pairwise correlation coefficients confirm the findings of previous analyses and show that the correlations observed between the measures are indeed statistically

significant. However, we should note that none of the correlations are not strong (but barely moderate), which indicates that the real dynamics of the Sözlük network depends also on other factors which were not addressed here.

The partial correlation analyses enabled us to test the different predictions of BA model and growth and deactivation model on the data.

The first finding is that when the effect of the current degree is removed no statistically significant correlation between the age and number of new links is observed

($r = -0.016$, $p > 0.450$). This finding is in accordance with the predictions of BA model. The age of a node does not seem to have an effect on acquiring new links (neither positive nor negative) independent of the current degree.

Secondly, we found out that when the effect of the recent degree increase is removed, the correlation between degree and number of new links decrease, but remain statistically significant ($r = 0.356$, $p < 0.01$). A similar but weaker effect is observed also in the partial correlation of recent degree increase and number of new links, controlled for node degree ($r = 0.132$, $p < 0.01$). Based on these results, we can claim that both recent degree increase and node degree has a moderate effect on the link acquisition, and this effect is independent to some degree because the remaining partial correlations are still significant.

Considering these findings, our conclusion is that both models have their own merits in predicting the growth of the network. In the short time scale, “Today’s Titles” facility and “today effect” is explained by the growth and deactivation model (i.e. power-law relation between age and likeliness to acquire new links) whereas in the longer time scales, the independency of age and new link acquisition probability can be explained by the BA model.

6 Conclusion

In this study, we introduced a new complex network data, which spans the whole lifetime of the network and contains very precise timestamps of the node and edge addition events. Using this dataset, we focused on the empirical validation of the linear dependency of preferential attachment on node degree. We introduced a new time independent measurement method which enabled us to independently analyze very short intervals and then integrating them altogether to provide measures for longer periods of time and capturing the dynamics more properly. The results suggest that the time independent method is more successful in extracting the linear relation between degree and preferential attachment for small degrees. The ability to analyze very short time periods also revealed an interesting relation (i.e. today effect) between the likeliness to acquire new links and age of a node, which was impossible to observe by the existing methods. As an application, the normalized measures were used to compare and evaluate two different network growth models on the present data.

References

1. Barabási, A.-L., Albert, R., Emergence of Scaling in Random Networks, *Science*, Vol. 286, pp. 509-512, 1999.
2. Newman, M.E.J., The structure and function of complex networks, *SIAM Review*, Vol. 45, pp. 167-256, 2003.
3. Pennock, D.M., Flake, G.W., Lawrence, S., Glover, E.J., Giles, C.L., Winners don't take all: Characterizing the competition for links on the web, *National Academy of Sciences*, Vol. 99, Issue 8, pp. 5207-5211, 2002.
4. Krapivsky, P.L., Redner, S., Leyvraz, F., Connectivity of Growing Random Networks, *Physical Review Letters*, Vol. 85, No. 21, pp. 4629-4632, 2000.
5. Klemm, K., Eguiluz, V.M., Highly clustered scale-free networks, *Physical Review E*, Vol. 65, 036123, 2002.
6. Vazquez, A., Growing Networks with Local Rules: Preferential Attachment, Clustering Hierarchy and Degree Correlations, *Physical Review E*, Vol. 67, 056104, 2003.
7. Dorogovtsev, S.N., Mendes, J.F.F., Evolution of Reference Networks With Aging, *Physical Review E*, Vol. 62, pp. 1842-1845, 2000.
8. Barabási, A.-L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., Vicsek, T., Evolution of the Social Network of Scientific Collaborations, *Physica A*, Vol. 311, pp. 590-614, 2002.
9. Eisenberg, E., Levanon, E.Y., Preferential Attachment in the Protein Network Evolution, *Physical Review Letters*, Vol. 91, 138701, 2003.
10. Jeong, H., Neda, Z., Barabási, A.-L., Measuring Preferential Attachment for Evolving Networks, *Europhysics Letters*, Vol. 61, pp. 567-572, 2003.
11. Newman, M.E.J., Clustering and Preferential Attachment in Growing Networks, *Physical Review E*, Vol. 64, No. 025102, 2001.
12. Redner, S., Citation statistics from more than a century of physical review, *physics/0407137*, 2004.
13. Eksi Sözlük, <http://www.eksisozluk.com>.
14. Rainbow, Rising, Polydor, 1976.

Table 1. Characteristics of Sözlük Data

Node Count	Link Count	Title Count	Entry Count	First Entry Date	Last Entry Date	Suser Count
1,543,328	5,016,632	1,543,328	2,414,296	15.02.1999	10.10.2004	29,712

Table 2. Pearson correlation coefficients of the measures, statistically significant correlations (i.e. $p < 0.01$) are printed in bold.

	$k_{t,i}^{norm}$	$\hat{l}_{t,i}^{norm}$	$a_{t,i}^{norm}$	$rdi_{t,i}^{norm}$
$k_{t,i}^{norm}$	-	-	-	-
$\hat{l}_{t,i}^{norm}$	0.427	-	-	-
$a_{t,i}^{norm}$	0.366	0.144	-	-
$rdi_{t,i}^{norm}$	0.406	0.283	0.088	-

Table 3. Partial correlation coefficients of the measures, statistically significant correlations (i.e. $p < 0.05$) are printed in bold.

Correlated Measures	Control measures	Partial Correlation Coefficients
$a_{t,i}^{norm}, \hat{l}_{t,i}^{norm}$	$k_{t,i}^{norm}$	-0.015
$k_{t,i}^{norm}, \hat{l}_{t,i}^{norm}$	$rdi_{t,i}^{norm}$	0.356
$rdi_{t,i}^{norm}, \hat{l}_{t,i}^{norm}$	$k_{t,i}^{norm}$	0.133

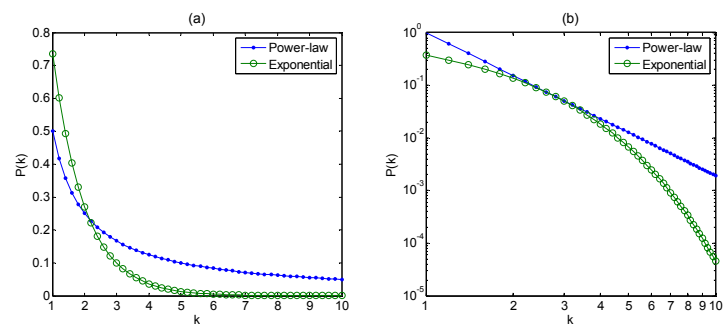


Figure 1. Power-law and exponential distributions: (a): Normal scale, (b): Log-log scale

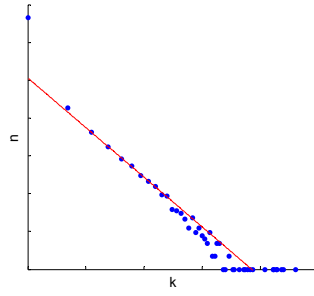


Figure 2. Log-log plot of the degree distribution of the resulting graph, and its best power-law fit

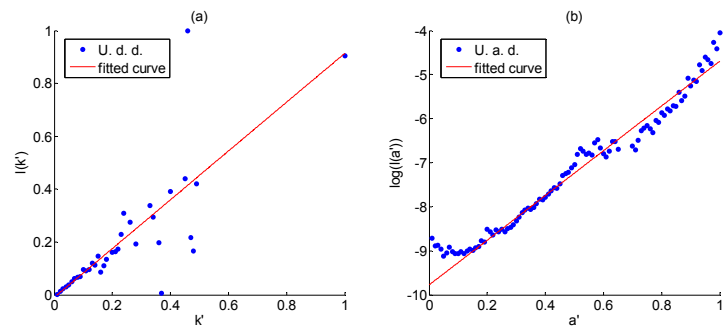


Figure 3. Number of new links as a function of (a): Degree, (b): Logarithm of age

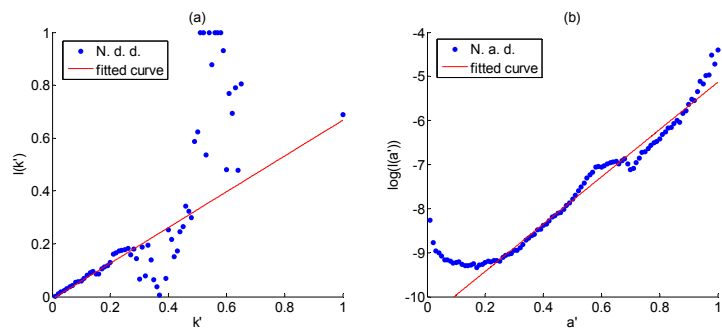


Figure 4. Number of new links as a normalized function of (a): Degree, (b): Logarithm of age