Soc. Netw. Anal. Min.
DOI 10.1007/s13278-011-0036-6

ORIGINAL ARTICLE

A social network evolution model based on seniority

Yi-Kuang Ko · Jing-Kai Lou · Cheng-Te Li ·
Shou-De Lin · Shyh-Kang Jeng

Received: 23 November 2010 / Revised: 6 July 2011 / Accepted: 1 August 2011
© Springer-Verlag 2011

Abstract A social network is a representation that describes the relationships between individuals. In recent years, several interesting phenomena, such as the densification power law, have been observed in social networks and a number of models have been proposed to explain them. In this paper, we investigate an interesting phenomenon called the seniority difference distribution of new connections, which we observed in our data. The distribution reveals that there exists a seniority preference when a new network node tries to establish connections with existing nodes. To explain the phenomenon, we propose several models based on different local selection policies, namely, equal-probability selection, freshness-based selection, oldness-based selection, and combined freshness/oldness selection. The results of simulations show that, by combining the concepts of freshness and oldness, it is possible to reproduce a social network that matches the observation.

Y.-K. Ko · J.-K. Lou · C.-T. Li · S.-D. Lin (✉) · S.-K. Jeng
National Taiwan University,
Taipei, Taiwan
e-mail: sdlin@csie.ntu.edu.tw

Y.-K. Ko
e-mail: r97921028@ntu.edu.tw

J.-K. Lou
e-mail: kaeaura@iis.sinica.edu.tw

C.-T. Li
e-mail: d98944005@csie.ntu.edu.tw

S.-K. Jeng
e-mail: skjeng@ew.ee.ntu.edu.tw

1 Introduction

1.1 Background

A social network generally consists of numerous nodes and edges, where each node represents an individual, and each edge represents a kind of relationship, such as kinship or friendship, between two nodes. By analyzing social networks, it is possible to observe and mine certain hidden knowledge or properties beyond the relationships to gain a deeper insight into the underlying structure of network.

In recent years, social network analysis (SNA) has become a popular interdisciplinary research area. The objectives can be divided into three tracks: measurement, modeling, and prediction.

1.1.1 Measurement

Researchers have identified certain interesting and important properties of social networks. For example, it has been observed that each node in a social network usually has a high clustering coefficient, which measures the compactness of a node's neighborhood (Watts and Strogatz 1998). It has also been observed that the diameter (average distances between pairs of nodes in a social network) of a social network is usually small. This is known as the 'small world' phenomenon (Watts and Strogatz 1998).

1.1.2 Modeling

Given certain phenomena observed in a social network (e.g.,small world), the next step is to design some computational or mathematical models to explain them. For example, two well-known topological features, the power-law degree distribution and the small world phenomenon,

Published online: 31 August 2011

are often observed in many real-world social networks. The power-law distribution states that the number of high-degree nodes varies as a power of the number of low-degree nodes (Newman 2003). The total number of k-degree nodes, n, can be expressed as a polynomial function $f(n) \propto n^{-k}$.

Several network generation models have been proposed to explain the above properties. The Barabasi–Albert (BA) model (Alber and Barabási 1999) and the Watts–Strogatz (WS) model (Watts and Strogatz 1998) are two famous examples. Barabasi et al. defined the preferential attachment process in which there is a higher probability that a new node will connect to higher-degree nodes in the network. It has been shown that the BA model can produce a network that follows the power-law distribution. To explain the small world phenomenon, Watts et al. proposed the WS model, which smoothly interpolates a random graph and a lattice by tuning a single rewiring parameter. The model demonstrates that a regular graph can be transformed into a small world network by rewiring a small proportion of edges at random.

1.1.3 Prediction

A more advanced exploration of social network data utilizes it for prediction; for example, to predict how a disease can spread in a given community, or how a network can grow or shrink over time. With the growing popularity of social network services, such as Facebook and Twitter, in recent years, how to explore social network data for advanced usage has generated a great deal of interest among researchers. For instance, in (Sakaki et al. 2010), the authors use social network data to predict the occurrence or trace the patterns of natural disasters such as earthquakes and typhoons.

In this paper, we focus on two objectives: (1) measurement and (2) modeling. Although many topological features of social networks have been observed and defined, there have been comparatively few attempts to investigate the evolution of a social network. Traditionally, a social network has been regarded as a static and cumulative summary of social activities. However, this perspective can be misleading in some situations. For example, it is assumed that high-degree nodes in a social network represent sociable individuals, but they could also represent capricious individuals who change friends constantly. Without knowing when the connections were established, it is hard to determine whether a high-degree node in a static network consistently possesses many neighbors or changes them along the time line.

To better understand social networks, researchers have analyzed dynamic social networks, which focus on studying the temporal information of nodes and edges. Dynamic social network analysis focuses on analyzing how static properties, such as degree, average path length, and clustering coefficient, evolve over time. For example, Leskovec et al. (2005b, 2007) proposed the densification power law and shrinking diameter, which are based on the evolutionary process of social networks. They found that when the number of nodes increases, a network becomes denser, and the graph's diameter decreases. Analysis of these properties reveals how the structure of a social network changes over time.

In this paper, we try to investigate the patterns of connection-search behaviors for those newcomers as they look for their neighbors. The preferential attachment property of new nodes has been known for several years. For example, under the BA model (Alber and Barabási 1999), a new node has higher chance of establishing connections with high-degree nodes. To gain further insight into the effect of the preferential attachment property, we also consider the temporal feature. Specifically, we consider the citation relations among papers as a kind of preferential attachment, and then study the evolution of citations in a citation network. Studying the effects of preferential attachment rule based on temporal information provides insights into the evolution of social networks and also benefits certain applications, such as prediction-based and recommendation-based social network services. Being able to predict which individuals are most likely to attract newcomers could yield a commercial advantage for social network services, because such information can be utilized to recommend friends or advertise merchandise.

1.2 Problem definition and proposed solution

In this work, we investigate under which temporal constraint the connections in a social network are created. We try to answer three questions:

(a) What kinds of temporal features can affect the decision of a new node about making new connections?
(b) Does a new node prefer to form relationships with well-established nodes or with other nodes that have joined the network recently?
(c) Is there a mathematical model to explain the answers to (a) and (b)?

First, we find that the attachment preferences of new nodes are affected by the age of existing nodes in the network. For instance, the interesting phenomenon can be found in US-Patent citation network that contains the patent citations in the USA between 1963 and 1999. The patents are the nodes and the citations are the edges in the US-Patent networks. As shown in Fig. 1, there is a bell-shaped distribution with a peak at the third year according

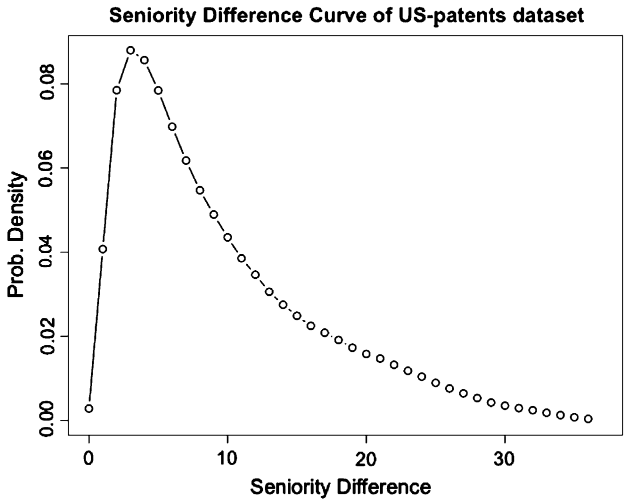

Fig. 1 Seniority difference distribution of the US-Patents dataset

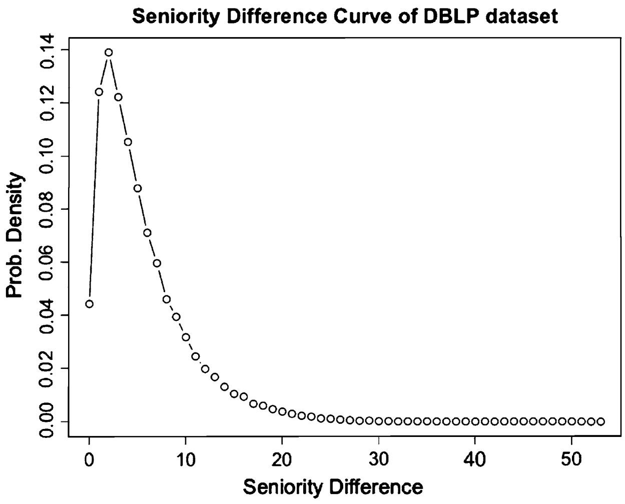

Fig. 2 Seniority difference distribution of the DBLP dataset

to the US-Patent citation network. Such distribution shows that new nodes attach to other nodes preferentially rather than at random. Generally, it is not easy for patents to attract a lot of attention as soon as they are published. On the other hand, out-of-date patents do not attract many citations either. Thus, there is a tendency for a newly published patent to cite other patents that are neither too old nor too new. To explain this finding, we develop a preferential selection model based on the seniority difference distribution of nodes. In the model, the probability that a new node will connect to an existing node is guided by their age difference. We believe that temporal preferential selection is driven by two criteria: the freshness and the oldness of existing nodes. To evaluate our idea, we conduct analytical proof as well as computer simulations on four mathematical models, to generate social networks of similar properties, and confirm that a model considering both freshness and oldness at the same time outperforms the others.

1.3 Contribution

The contribution of this paper is twofold: (1) we identify an interesting phenomenon in the data, namely, there is a different kind of preferential attachment behavior in social networks. Because new nodes tend to connect to existing nodes that joined the network several time-steps away, we call this behavior temporal preferential attachment distribution (or seniority difference distribution), as shown in Figs. 2, 3, and 4. To the best of our knowledge, this is the first work to discuss the relationship between the difference in seniority and preferential attachment. (2) We propose a graph-generation model to explain the temporal preferential attachment distribution. The model exploits the idea of freshness and oldness. We believe that a node possessing certain levels of both properties will have a higher probability of being connected. Our simulation results show that the model can reproduce networks that match the observed temporal preferential attachment distribution of new nodes.

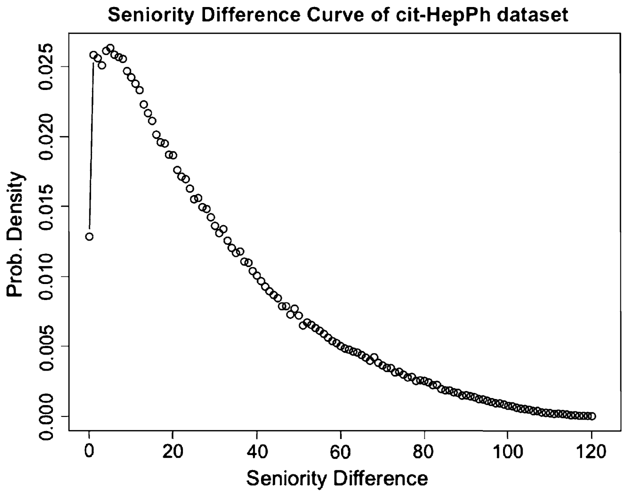

Fig. 3 Seniority difference distribution of the cit-HepPh

The remainder of the paper is organized as follows. The next section provides a review of related works. In Sect. 3, we show our observation from data. In Sect. 4, we provide the details of the model and the mathematical analysis. In Sect. 5, we list the experiment results from a different setup. Section 6 contains some concluding remarks.

2 Related work

Research on social network generation focuses on designing models that can generate graphs with similar properties

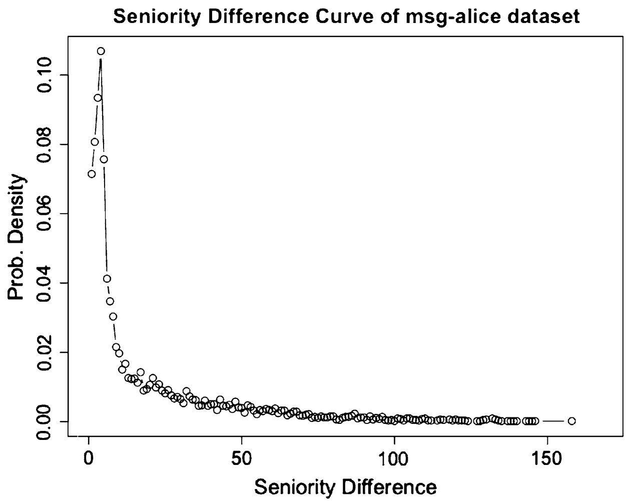

Fig. 4 Seniority difference distribution of the Fairyland Online

to those observed in real-world social networks. Table 1 lists several well-known models and the properties they try to explain.

2.1 Erdős-Rényi (ER) model

The ER model (Erdős and Rényi 1959) is a fundamental graph-generation model that constructs graphs by fixing the number of nodes and assuming there is a certain probability p to create each edge between all possible pair of nodes. It can be proved that ER model can only produce graphs that follow a Poisson degree distribution rather than a power-law distribution. However, by adjusting the given probability for generating edges between node pairs, it is possible to discover the phase transition phenomena for a large component. Although the ER model does not fit real-world phenomena perfectly, it is the basis of many network generation models.

2.2 Configuration model

The configuration model tries to generate graphs whose degree distribution is similar to a given degree sequence. However, the generated graph cannot satisfy other properties, such as a high clustering coefficient or a small diameter. Mahadevan et al. (2006) extend the configuration model by proposing the dK-graph to fit higher-degree distributions. The goal is to produce a graph that is even closer to the real graph.

2.3 Barabasi–Albert (BA) model

The BA model utilizes “the rich get richer” concept to realize the power-law distribution property. By assuming that nodes with a higher degree have a higher probability of being connected, the BA model can produce a network with a power-law distribution $p(k) \propto k^{-3}$, where k is the degree (Bu and Towsley 2002).

2.4 Watts–Strogatz (WS) model

The WS model tries to explain the small world theory, which states that “the average pair-wise distance between nodes is small in a social network”. The model interlaces two types of graph structure, a lattice graph and a random graph, by utilizing the rewiring probability in the lattice to reconnect a node to longer-distance nodes. It has been shown that, even with a small rewiring probability, it is

Table 1 Models and properties

Model name	Properties
ER model (Erdős and Rényi 1959)	Poisson distribution, phase transition
Configuration model (Newman 2003)	Given a certain degree distribution to generate a network with that distribution
dK-graphs (Mahadevan et al. 2006)	K-joint degree distribution
BA model (Alber and Barabási 1999)	Power-law degree distribution, preferential attachment
WS model (Watts and Strogatz 1998; Kleinberg et al. 1999)	Small world properties: high clustering coefficient, low average path length
Winners don’t take all (Pennock 2001)	Log-normal distribution with different modes
Random walk and nearest neighbor Vázquez (2003)	Power-law degree distribution, clustering distribution, average nearest neighbor degree distribution
Forest fire model (Leskovec et al. 2005b, 2007; Pedarsani et al. 2008)	Densification law, shrinking diameter, power-law distribution
Utility-based model (Du et al. 2009)	Clique degree power law, triangle weight power law, clique participant power law
Recursive tensor model (Akoglu and Mcglohon 2008; Leskovec and Faloutsos 2007; Leskovec et al. 2005a; Kossinets and Watts 2006)	Densification law, weighted power law, λ_1 power law, $\lambda_{1,w}$ power law, edge weight power law

possible to create a network with the small world property.

The above four models are the most commonly known models trying to fit the well-known properties in real social networks. Among them, only the BA model considers the evolutionary process of a network. The following advanced models are based on the BA model.

2.5 Winners don't take all model

Pennock et al. (2001) re-examined the power-law distribution property and found that if the Internet is divided into fine-grained networks based on sub-categories, such as companies or universities, it is possible to find a log-normal degree distribution among the nodes; however, the higher-degree nodes still follow a power-law distribution. They then extended the BA model to explain the log-normal distribution.

2.6 Random walk model and nearest neighbor model

Vázquez (2003) showed that the power-law distribution property could be achieved by using part of the social network rather than the whole network. In other words, when a node joins a network, it does not need to know the degree of each node in the graph; instead, it only needs the information about the subset of nodes to be targeted. The authors propose two models based on local rules, a random walk model and a nearest neighbor model, to reproduce the power-law distribution.

2.7 Forest fire model

Leskovec et al. (2005b, 2007) proposed the forest fire model to explain two observed properties: the "densification power law" and "shrinking diameter" properties of dynamic graph evolution. Previous works assumed that the number of edges grows linearly with the number of nodes. However, after investigating certain real-world networks, the authors obtained a different result; namely, the number of edges grows super-linearly with the number of nodes. This implies that a social network becomes denser over time with the average degree increasing with its size. The authors also found that the effective diameter decreases when new nodes are added to the network and they proposed the forest fire model to produce graphs that possess these two properties.

When a node joins a graph under the forest fire model, it first selects and connects to one or more existing nodes (denoted as ambassadors). Then, it recursively selects the neighborhood nodes of its connected nodes (based on a forward selection probability P_a and a backward selection probability P_b) to establish further connections, just like a fire spreads in the forest. A literature survey provides a good analogy to this model; for example, a researcher starts by reading a relevant paper, which cites other papers for the researcher to read (connect to). Then they show that the model can reproduce a network with the two observed properties via simulations.

2.8 Utility-based model

Du et al. (2009) observed some clique-based properties in a personal communication social network. They found that the number of cliques that each node belongs to forms a power-law distribution with respect to the node degree, and the size of the clique follows the power law. To explain this finding, they proposed a utility-based model in which the agents join the network gradually and try to optimize a certain utility function. The observed behavior occurs when each agent tries to optimize the given utility function.

2.9 Recursive tensor model

Akoglu and Mcglohon (2008), Leskovec and Faloutsos (2007), Chakrabarti et al. (2004), and Leskovec et al. (2005b), exploited the idea of the entropy plot to discover the structure's fractal patterns during a graph's evolution. The authors propose to use a three-dimensional tensor to represent a graph by adding a time dimension, and combine Zipf's law and two-dimensional random typing to produce graphs that fit a list of observed properties.

As far as we know, this paper is the first work that talks about such seniority difference and proposes a model to explain it. Next, we describe a novel phenomenon that we observed in our data and propose a mathematical model to explain it.

3 The temporal preferential attachment property

First, we describe the datasets used in our simulations and then discuss the observed temporal preferential attachment phenomenon.

3.1 Dataset description

We used three five datasets in the analysis: the cit-HepPh dataset (KDD Cup 2003. http://www.cs.cornell.edu/projects/kddcup/datasets.html), the US-Patent dataset (National Bureau of Economic Research. http://www.nber.org/), the DBLP dataset (http://www.informatik.uni-trier.de/~ley/db/), the Fairyland Online dataset (http://fairyland.lagernet.com/), and the Plurk dataset (http://www.plurk.com/).

1. cit-HepPh dataset: The dataset was used in the ACM KDD (Cup 2003) and contains the citation information of papers published in the field of high energy physics from January 1993 to April 2003, a total of 124 months. For each paper, the dataset records the authors' names, the abstract, the references, and the year of publication; thus, it is possible to learn whether a paper tends to cite old papers or new ones. Based on the cit-HepPh dataset, we construct a simple directed graph called a cit-HepPh citation network to represent the citation relations among the papers in the high energy physics field. The nodes represent the published papers and the directed edges are drawn from the papers that they cite. There are 34,546 nodes (published papers) and 421,578 edges (citations) in the cit-HepPh citation network.
2. US-Patent dataset: The dataset, which was compiled by the National Bureau of Economic Research, contains 3,923,922 patents published in the USA between 1963 and 1999. Based on the dataset, we can construct a US-Patent citation network in the same way that we construct the cit-HepPh citation network. However, as the dataset does not contain citation information for patents published before 1975, the constructed citation network is only based on patents published from 1975 to 1999. The network comprises 3,774,362 nodes (patents) and 16,512,783 edges (citations).
3. DBLP dataset: The DBLP dataset is a computer science bibliography Web site that lists more than 1.3 million papers. We extract papers published in the period from 1945 to 2006 with the citation information to construct the DBLP citation network, which contains 122,995 nodes. Roughly, 25.8% of the referenced publications are not available in the dataset.
4. Fairyland Online dataset: Fairyland Online is a massively multiplayer online role-playing game, which was developed by Lager Network Technologies and operates in Taiwan, Hong Kong, Mainland China, Thailand, and South Korea. We collected the chat logs from a server of Fairyland Online called Alice starting from February 2003 to April 2004. This dataset contains 32,690 player IDs and 15,789,502 messages. An edge is created when there is a chat behavior between users. The message networks consist of 32,690 nodes and 445,528 edges.
5. Plurk message dataset: Plurk is a well-known Microblog Web site which allows people to post messages and reply to existing posts. We crawl the Plurk messages since 1 February 2009 to 24 May 2009, and 249,508 messages are collected. We create a Plurk message network with 248,115 nodes and 2,709,720 edges, where the nodes stand for user IDs and edges stand for the post-reply relationship among users.

For the first three datasets (cit-HepPh, US patent, and DBLP), the joining time for each node in the network is exactly given. For the later two datasets (Fairyland Online and Plurk), the joining time for each user is not available. Here, we assume the joining time of each node is uniformly distributed among a small span that precedes its first chat time.

3.2 Seniority difference distribution

In this sub-section, we consider two questions: (1) when a node (i.e., an individual) joins a network or a community, does the seniority of the target nodes play an important role in its choice of neighbors? If the answer is yes, (2) does the node tend to establish connections with the most senior or the most junior members, or neither type? We found that the seniority of the target nodes does play an important role in the establishment of connections, and a new node prefers not to connect to the youngest or oldest nodes.

The seniority of a node in a dynamic social network is defined as the span of its existence in the network. The seniority difference of an edge is defined as the absolute difference between the seniority values of two end points. For instance, in a citation network, the seniority difference of an edge represents the time difference between when a paper was published and when it was cited by another paper. To gain insight into the preferential strategy based on seniority, we investigate the seniority difference distribution, i.e., the distribution of the seniority difference of all edges in the network.

We assume the dynamic social network structure in studying the attaching process of new nodes over time. Specifically, we assume new nodes/edges are added to a social network continuously. This is normal since people join a network continuously, and papers or patents are published continuously. A dynamic social network can be represented as a simple directed graph, G (V, E), where V is the set of nodes, and E is the set of edges. The time stamp of when a node v joins the network is denoted as t_v and the time stamp of when a link e first appears in G is denoted as t_e. The seniority value of each new node is set at zero and then increased gradually over time.

Here, a new node with respect to time t is defined as a node that joins the network no later than time $t + \varepsilon$, where ε is a small value that indicates a grace period. The nodes are considered to be fresh during the grace period, and we are only interested in the seniority difference of the linked node pairs while the attaching nodes are fresh.

We set the grace period ε to 1 (day) for the message networks in this work, and $\varepsilon = 0$ for citation networks. The detailed method for generating such a distribution is described below.

Algorithm: Seniority Difference Distribution for New Nodes

Input: a dynamic social network G(V, E), a grace period ε

Output: the seniority difference distribution F

```
For each $e = v \rightarrow w$ in G(V, E) do:
    if $t_e - t_v \leq \varepsilon$ then:
        $F(t_v - t_w) = F(t_v - t_w)+1$
    end if
end for
Normalize F as a distribution
```

Note that there are a few negative seniority difference values in the cit-HepPh dataset (i.e., a paper cites another paper that was published later) and the US-Patent dataset. We simply ignore those cases while generating the distribution.

We compute the seniority difference distribution on the three datasets and show the distribution in Figs. 1, 2, 3, 4, 5. Interestingly, there is generally one major hump in the distribution that lies slightly deviated from the origin. It shows that new nodes tend to connect to young nodes, but not the youngest. Another interesting observation is that the right-hand side of the distribution does not seem to follow a power-law distribution. The drop is not very significant and there is no long tail. This shows that new nodes do connect to senior nodes in certain circumstances, possibly when the senior nodes represent a small number of classical papers or patents. Generally, the distribution reveals that new nodes select targets preferentially based on seniority.

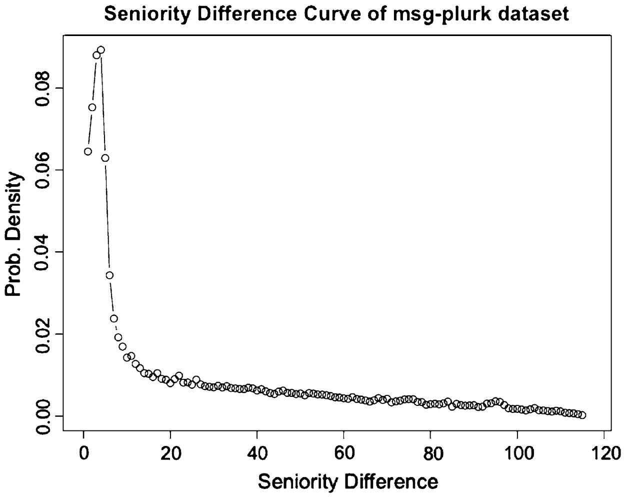

Fig. 5 Seniority difference distribution of the Plurk

4 Local preferential selection model and analytical results

4.1 Preferential selection model

As mentioned in Sect. 3, social networks have a temporal preferential attachment property. Since the property describes a global distribution over all edges, we would like to determine whether there is some local preference allowing new nodes to follow such a global distribution.

To maintain the densification power law and shrinking diameter properties, the design of our model is based on modifying the forest fire model, which gradually evolves graphs via a random walk process. Nodes arrive one at a time and form outlinks to some subset of the earlier nodes by attaching it to a random node w called ambassador in the existing graph, and then begins establishing links outward from w, with a certain probability recursively.

In contrast to the forest fire model, rather than choosing the ambassadors randomly, our model uses the following four local seniority-based strategies to select nodes: equal-probability selection, freshness-based selection, oldness-based selection, and hybrid selection.

We perform two kinds of analysis on the models: when the fire propagation probability of ambassadors is negligible (i.e., the forward probability $P_a =$ the backward probability $P_b \approx 0$), it is possible to derive analytical solutions for the seniority distribution. We will discuss this aspect later in this section. When the fire propagation probabilities are not negligible, no closed-form solutions exist; therefore, we use computer simulations to verify the fit. We discuss the results in the next section.

4.1.1 Equal-probability seniority selection

This strategy assumes each senior node has an equal probability of being chosen. The selection strategy is treated as a baseline scheme for comparison. Basically, equal-probability seniority selection stands for non-preferential selection in the temporal aspect.

Figure 6 shows the seniority difference distribution of both the US-Patent dataset and the one generated with equal-probability selection through simulation. Clearly, the equal-probability strategy fails to produce a network with the real-world temporal preferential attachment property; thus, it is reasonable to conclude that the new edges are not formed at random in terms of seniority. Although the equal-probability selection stands for non-preferential in the temporal aspect, the senior nodes will eventually receive more connections since they stay in the network for a longer period of time. Hence the equal-probability selection would not result in the uniform distribution.

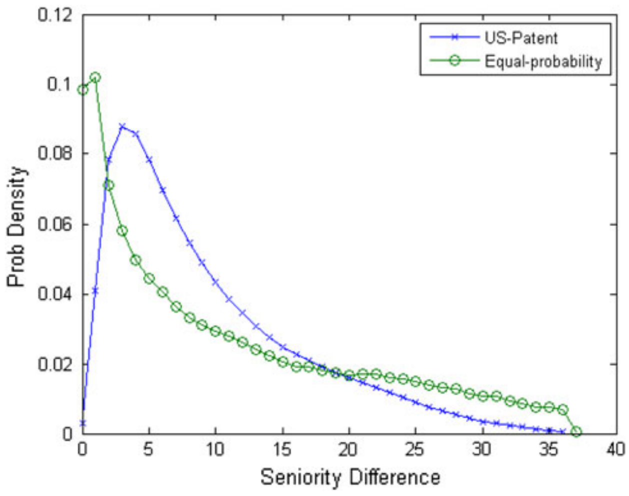

Fig. 6 The seniority difference curve of the US-Patent dataset derived by the equal-probability local selection method (assuming $P_\mathrm{f} = 0.3$, $P_\mathrm{b} = 0.3$) (see the comment on the right)

Given negligible P_a and P_b in the forest fire model, it is possible to derive an analytic solution for the model to show that the strategy cannot fit the observed curve. Let the observation period of the dynamic social network be the interval of $[0, t_\mathrm{f}]$. Also let $S(d)$ be the total number of edges of seniority difference d, and $S(d, t)$ be the number of edges established at time t that have seniority difference d. Since the nodes that joined before time d cannot establish edges with a seniority difference larger than or equal to d, we can conclude that

$$S(d) = S(d, d+1) + S(d, d+2) + \cdots + S(d, t_f) \tag{1}$$

Furthermore, since the seniority difference values of edges established at time t range from 1 to t (assuming a node does not attach to nodes of the same age), and since the local seniority difference is distributed equally over time, we can infer that

$$S(d, t) = \frac{N_t}{t}, \tag{2}$$

where N_t is the total number of edges established at time t.

By combining Eqs. 1 and 2, we obtain

$$S(d) = \frac{N_{d+1}}{d+1} + \frac{N_{d+2}}{d+2} + \cdots + \frac{N_{d+t_f}}{d+t_f}.$$

If we assume that P_a and P_b are close to 0, we can infer that an equal number of edges are established in each time stamp (i.e., $N_t = N_{t-1} = N_{t+1}$ for all t). Then, we can obtain the distribution of $S(d)$ as follows:

$$S(d-1) : S(d) : S(t_f) = \left(\frac{1}{d} + \cdots + \frac{1}{t_f}\right)$$
$$: \left(\frac{1}{d+1} + \cdots + \frac{1}{t_f}\right) : \frac{1}{t_f}.$$

We show the seniority distribution as the line with circle dots in Fig. 9, which significantly diverges from the true observed distribution.

4.1.2 Freshness-based selection

We assume that new nodes prefer to attach to young nodes in the network. This is reasonable as new members can easily establish friendships with other young members since they are similar to each other in seniority. In a dynamic social network, this property can be satisfied by providing higher probabilities to fresh nodes while trying to establish a link. We assume that the probability of establishing links based on seniority values follows the power-law distribution $p(s) \propto s^{-k}$, where s is the seniority difference value and k is a parameter that controls the drop rate of the curve. Note that we can also use other kinds of distributions such as exponential distribution and linear-line distribution of positive slope to obtain similar results.

The simulation results of particular settings are shown by the square-dot curve in Fig. 7. Clearly, the strategy fails to explain the observed phenomenon. Similar to the previous strategy, we can derive the distribution of the selection strategy given negligible propagation probabilities. Note that Eq. 1 still holds in this case, but Eq. 2 must be changed to

$$S(d, t) = \frac{N_t d^{-k}}{Z_t}, \text{where the normalization factor } Z_t$$
$$= \sum_{x=1}^{t} x^{-k}$$

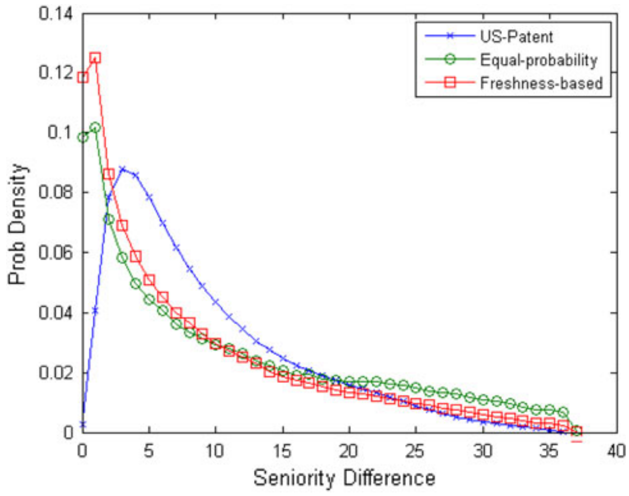

Fig. 7 The seniority difference curve of US-Patent dataset derived by the freshness-based local selection method (assuming $k = 1.5$ and $P_\mathrm{f} = 0.3$, $P_\mathrm{b} = 0.3$)

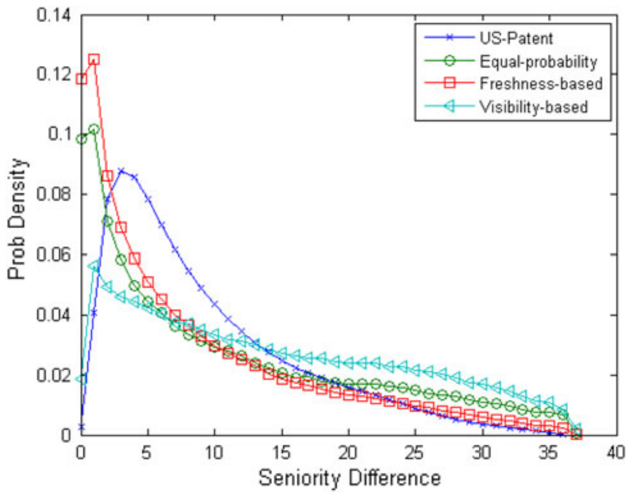

Fig. 8 The seniority difference curve of US-Patent dataset derived by the oldness-based local selecting method (assuming $k = 1.5$ and $P_\mathrm{f} = 0.3$, $P_\mathrm{b} = 0.3$)

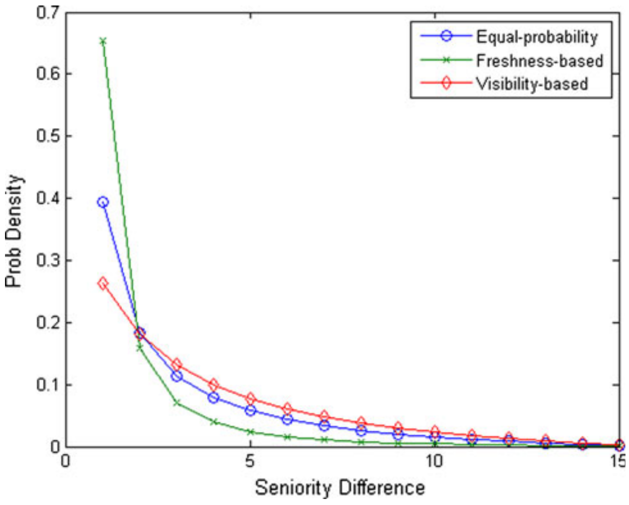

Fig. 9 The analytic solution for the seniority difference curves derived by for the three compared selection methods with $P_\mathrm{f} = 0$, $P_\mathrm{b} = 0$

Therefore, we can derive

$$S(d) = \frac{N_t d^{-k}}{Z_{d+1}} + \frac{N_{t+1} d^{-k}}{Z_{d+2}} + \cdots + \frac{N_{t_f} d^{-k}}{Z_{t_f}}$$

$S(d) = Nd^{-k} \sum_{i=d+1}^{t_f} \frac{1}{Z_i}$, where i in $[1, t_\mathrm{f}]$

Assuming that $P_\mathrm{a} = P_\mathrm{b} \approx 0$ implies that N_t is a constant over t, we can obtain

$$S(d-1) : S(d) : S(t_f) = \left(\frac{(d-1)^{-k}}{Z_d} + \cdots + \frac{(d-1)^{-k}}{Z_{t_f}}\right) : \left(\frac{d^{-k}}{Z_{d+1}} + \cdots + \frac{d^{-k}}{Z_{t_f}}\right) : \frac{t_f^{-k}}{Z_{t_f}}$$

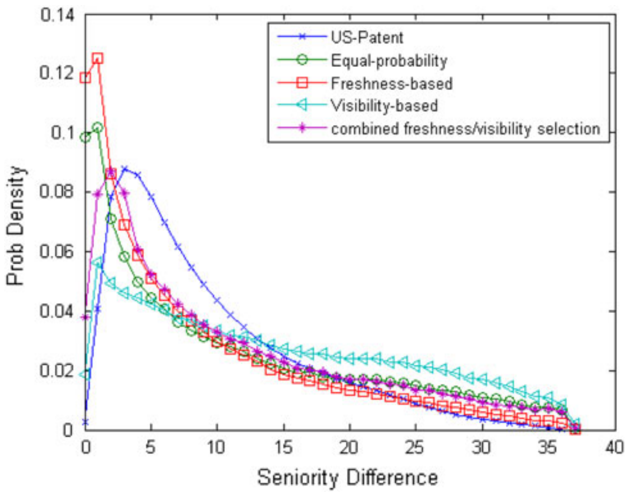

Fig. 10 The seniority difference curve of the US-Patent dataset derived by the hybrid local selection method. The mean, left sigma, and right sigma parameters equal 3, 1, and 20 respectively (assuming $P_\mathrm{f} = 0.3$, $P_\mathrm{b} = 0.3$)

The result of the above distribution is also shown in Fig. 9. The results confirm that the strategy is not the best fit for the data.

4.1.3 *Oldness-based selection*

Differing from freshness-based selection, the oldness-based selection strategy tends to connect new nodes to more senior nodes. We assume that higher seniority implies higher oldness. The oldness-based strategy tries to capture the fact that a new member prefers to establish connections with senior members to consolidate its own position. Here, we also model the connection probability using power-law form $p(d) \propto d^k$, where d is the seniority difference and k is a parameter that controls the slope.

Figure 8 shows how the strategy affects the global seniority difference distribution. In this model, the left part of the curve is not as steep because new nodes prefer to connect to nodes that have a large seniority difference. However, the low seniority difference cases still possess the higher probability because of the fact that higher seniority difference cannot occur until the later stage. Using similar derivation process to that discussed previously, it is possible to derive an analytic solution for the seniority distribution as follows:

$$S(d-1) : S(d) : S(t_f) = \left(\frac{(d-1)^k}{Z_d} + \cdots + \frac{(d-1)^k}{Z_{t_f}}\right) : \left(\frac{d^k}{Z_{d+1}} + \cdots + \frac{d^k}{Z_{t_f}}\right) : \frac{t_f^k}{Z_{t_f}},$$

where $Z_t = \sum_{x=1}^{i} x^k$.

The distribution of the above solution is shown in Fig. 9. Clearly, the model cannot reproduce the observed phenomenon either.

4.1.4 *Combined freshness/oldness selection*

Although the previous two models are intuitive and reasonable, neither of them can faithfully explain our observation. Therefore, to allow the freshness and oldness properties, we combine the concepts with the selection process. For example, a new member might prefer to know somebody that is not too far away in terms of age (to avoid the generation gap), but still senior enough to be able to teach the new member. Therefore, we propose a half-mix Gaussian model to form a hybrid of the freshness-based selection and oldness-based selection strategies. The proposed model can express two types of selection behavior simultaneously. We use a mean-shift half-Gaussian distribution model as the local selection distribution: the left half-Gaussian models freshness preferences and the right half-Gaussian models oldness preferences. There are three parameters in the model: the peak that separates these two distributions, the right variance and the left variance. As shown in Fig. 10, the model can reproduce a seniority difference distribution successfully.

As shown in Fig. 10, the hybrid model fits the observation better than three previous models. The graph-generation pseudo code based on different local selection strategies is as follows:

Algorithm Preferential Selection Model (modified from Forest Fire Model)

Inputs: Graph at time t-1, G_{t-1}, newly joined nodes $\{v_1, \ldots, v_n\}$, the forward probability p_f , and the backward probability ratio p_b; x is a random variable of the geometric distribution with mean $p_f/(1-p_f)$, and y is a random variable of the geometric distribution with mean $p_b/(1-p_b)$

Output: Graph at time t, G_t

```
1. For $v_i$ in newly joined nodes $\{v_{1,} v_2, \ldots, v_n\}$ do
2.     For any w in V ($G_{t-1}$), $p_w$ = SelectingFunction($t_v$, $t_w$)
3.         $v_i$ selects the first node m as the ambassador node with probability $p_w$
4.       Generate x, the number of in-link neighbors that ambassador node need to select
5.       Generate y, the number of out-link neighbors that ambassador node need to select
6.       Select x in-link neighbors of $v_i$ , $v_{i,1} \ldots v_{i,\,x}$ based on
          the connecting probabilities of these in-links
7.       Select y out-link neighbors of $v_i$, $v_{i,x+1}$,$v_{i,x+y}$ based on
          the connecting probabilities of these out-links
8. Perform Steps 2-4 for each node of $v_{i,1}$ to $v_{i,\,x+y}$. As the process continues, nodes cannot be
   revisited in order to prevent the construction from cycling
9. Build links from $v_i$ to nodes that $v_i$ has visited during the process in Steps 2-9.
```

Procedure Selection Function (t_v, t_w)

Input: time stamp of a node v and an earlier node w, the parameters for each selection method

Output: the connecting probability between v and w, denoted as $p_{w,c}$

```
1. If the selection method is "Equal-probability selection" return 0.5
2. If the selection method is "Freshness-based selection" return $(t_v-t_w)^{-k}$
3. If the selection method is "Oldness-based selection" return $(t_v-t_w)^{k}$
4. If the selection method is "Combined Freshness/oldness selection", then
5.     If $t_v-t_w$ <= mean of the mixed distribution, then return N($t_v-t_w$-mean, left sigma)
       else return N($t_v-t_w$-mean, right sigma)
```

Table 2 KL divergence of the US-Patent dataset

Function/parameter	$P_f = 0$, $P_b = 0$	$P_f = 0.3$, $P_b = 0.2$	$P_f = 0.3$, $P_b = 0.3$
Equal-probability selection	0.0846224	0.341618	0.269289
Freshness-based selection	0.382915	0.352888	0.387504
	$K = 1.5$	$K = 2$	$K = 1.5$
Oldness-based selection	0.329092	0.426805	0.302748
	$K = 1.5$	$K = 1.5$	$K = 1.5$
Combined freshness/oldness	0.00728241	0.111592	0.111122
selection (mean, variance_left, variance_right)	(3, 1, 20)	(3, 1, 5)	(3, 1, 10)

5 Simulation results

5.1 Seniority difference curve

In the discussion of the previous section, we assumed that the total number of new links added at every time stamp is identical; however, it seems that this is not the case in the real world. Therefore, we relax the constraint when performing simulations. For each dataset, we determine the number of nodes added to the network at every time stamp from real data, and use the information in the simulations. We then apply the four proposed selection strategies for network generation. To measure the fit of each model to the observed distribution, we use the Kullback–Leibler distance (i.e., KL divergence) as the performance metric. The KL divergence of two distributions, P and Q, is defined as

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

However, KL divergence is not a symmetric metric, which means $D_{\mathrm{KL}}(P\|Q)$ is not equal to $D_{\mathrm{KL}}(Q\|P)$. To resolve this problem, researchers usually exploit the symmetric KL divergence defined as

$$D_{\mathrm{KL,Symmetric}}(P \parallel Q) = D_{\mathrm{KL}}(P \parallel Q) + D_{\mathrm{KL}}(Q \parallel P)$$

The smaller the value, the better one distribution will match another one.

Table 3 KL divergence of the cit-HepPh dataset

Function/parameter	$P_f = 0$, $P_b = 0$	$P_f = 0.3$, $P_b = 0.2$	$P_f = 0.3$, $P_b = 0.3$
Equal-probability selection	0.013695	0.208685	0.312501
Freshness-based selection	0.228278	0.143297	0.129021
	$K = 1.5$	$K = 3$	$K = 1.5$
Oldness-based selection	0.474986	0.467015	0.344167
	$K = 1.5$	$K = 1.5$	$K = 1.5$
Hybrid freshness/oldness	0.611096	0.038217	0.066905
selection (mean, variance_left, variance_right)	(5, 3, 20)	(5, 1, 20)	(5, 1, 20)

The symmetric KL distances between the seniority differences of the observed networks and those of the networks generated by our models are listed in Tables 2, 3, and 4; and the corresponding distribution curves are shown in Figs. 11, 12, and 13. Overall, the hybrid model captures the shape of the distribution more accurately than the other models and its KL values are significantly better (Figs. 14, 15; Table 5).

Function/parameter	$P_f = 0$, $P_b = 0$	$P_f = 0.3$, $P_b = 0.2$	$P_f = 0.3$, $P_b = 0.3$
Equal-probability selection	0.56468	0.868125	1.302026
Freshness-based selection	0.196089	0.476171	0.259168
	$K = 2.5$	$K = 3$	$K = 3$
Oldness-based selection	1.193813	1.302026	1.126704
	$K = 1.5$	$K = 1.5$	$K = 1.5$
Hybrid freshness/oldness	0.447574	0.165145	0.232614
selection (mean, variance_left, variance_right)	(3, 20, 20)	(3, 1, 3)	(3, 1, 3)

Function/parameter	$P_f = 0$, $P_b = 0$	$P_f = 0.3$, $P_b = 0.2$	$P_f = 0.3$, $P_b = 0.3$
Equal-probability selection	0.229486	0.228846	0.191136
Freshness-based selection	0.551131	0.120874	0.149213
	$K = 2$	$K = 2$	$K = 1.5$
Oldness-based selection	0.378009	0.444451	0.404766
	$K = 1.5$	$K = 1.5$	$K = 1.5$
Hybrid freshness/oldness	0.214659	0.13175	0.101619
selection (mean, variance_left, variance_right)	(3, 1, 50)	(3, 1, 20)	(3, 1, 50)

5.2 Discussion

We performed another simulation to determine whether the long-term, global seniority difference distribution could be modeled by using the identical local seniority difference distribution of a given time stamp. Specifically, we assume that, for each time stamp, there exists an oracle revealing the overall long-term seniority difference, and the nodes can use such distribution to select the nodes to connect. For example, in the simulation on the US-Patent dataset, every new node uses the distribution shown in Fig. 2 to establish links. At the first glance, this strategy should produce a distribution that matches the global and long-term distribution. However, the simulation results show that the strategy is not as effective as the hybrid strategy.

We believe the reason is that, for any given time stamp t, edges can only be created if the seniority difference is smaller than t. The global distribution can be regarded as an accumulation of a series of edge differences within different

Table 4 KL divergence of the DBLP dataset

Function/parameter	$P_f = 0$, $P_b = 0$	$P_f = 0.3$, $P_b = 0.2$	$P_f = 0.3$, $P_b = 0.3$
Equal-probability selection	0.169237	0.997186	0.555309
Freshness-based selection	0.038338	0.639090	0.329910
	$K = 1.5$	$K = 3$	$K = 2.5$
Oldness-based selection	1.497535	1.579189	1.144650
	$K = 1.5$	$K = 1.5$	$K = 1.5$
Hybrid freshness/oldness selection (mean, variance_left, variance_right)	0.031695	0.209268	0.126830
	(2, 3, 10)	(2, 20, 3)	(2, 1, 3)

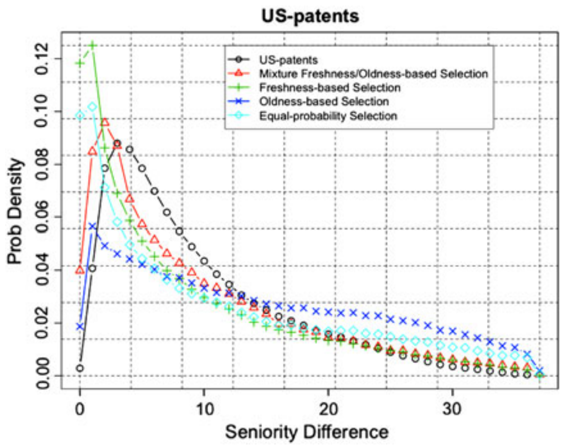

Fig. 11 Comparison of the seniority difference distribution of different selection methods using the US-Patent dataset ($P_f = 0.3$, $P_b = 0.3$)

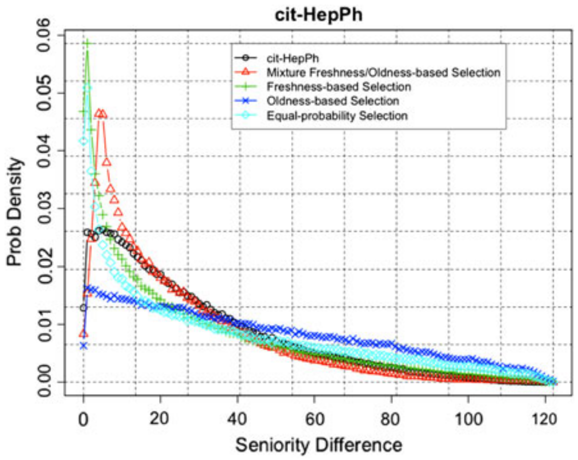

Fig. 12 Comparison of the seniority difference distribution of different selection methods using the cit-HepPh dataset ($P_f = 0.3$, $P_b = 0.3$)

time stamps. Because the sample space of each time period is different, there is no reason the local distribution should match the global and accumulated distribution.

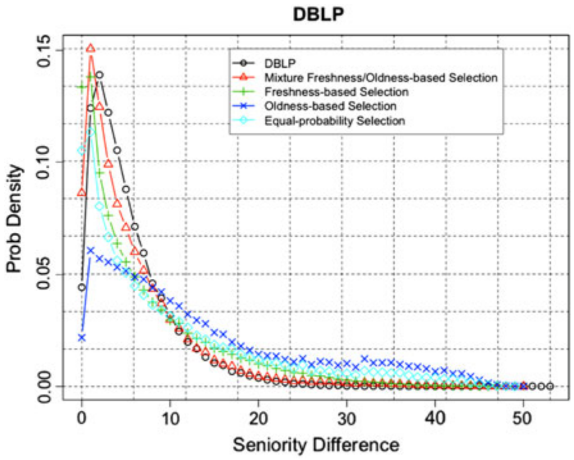

Fig. 13 Comparison of the seniority difference distribution of different selection methods and the DBLP dataset ($P_f = 0.3$, $P_b = 0.3$)

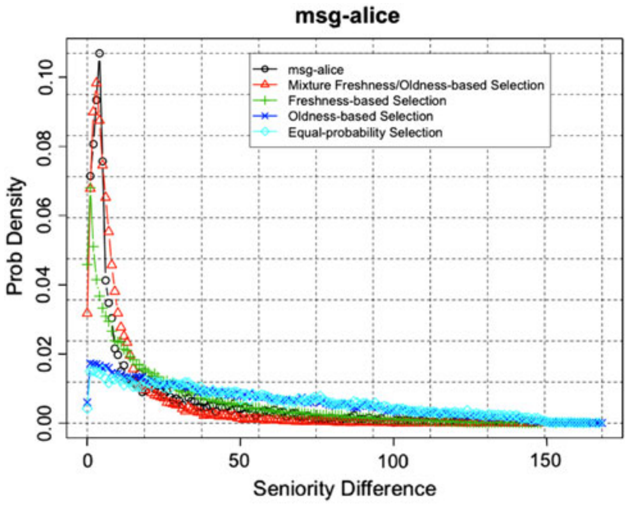

Fig. 14 Comparison of the seniority difference distribution of different selection methods and the Fairyland Online dataset ($P_f = 0.3$, $P_b = 0.3$)

6 Conclusion

Researchers have long been interested in investigating how connections in a social network are established, and several attachment preferences have been confirmed previously. This work proposes another kind of preferential attachment strategy in the temporal dimension, namely, the seniority difference distribution. The statistics derived from real data show that the seniority difference distribution is bell shaped, and the peak is slightly deviated from the origin, while the tail does not follow a power-law distribution. We propose several intuitive selection strategies to explain such an observed phenomenon. Finally, we show that by

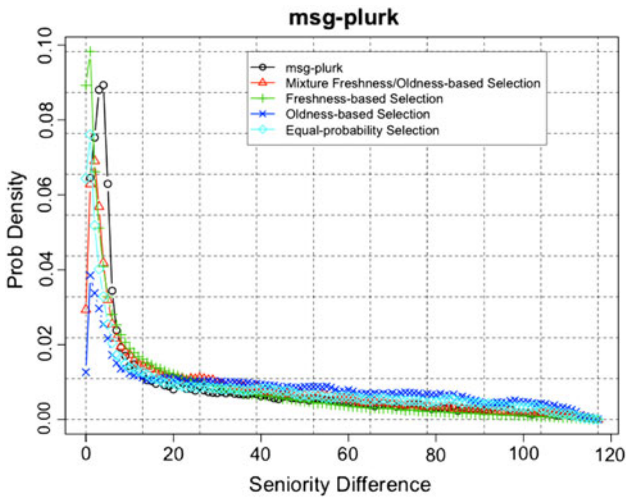

Fig. 15 Comparison of the seniority difference distribution of different selection methods and the Plurk dataset ($P_f = 0.3$, $P_b = 0.3$)

Table 5 KL divergence of the US-Patent dataset, using the observed global distribution as the local selection probability

Function/parameter	$P_f = 0$, $P_b = 0$	$P_f = 0.25$, $P_b = 0.25$	$P_f = 0.3$, $P_b = 0.4$	$P_f = 0.5$, $P_b = 0.5$
Hybrid freshness/ oldness	0.0233324	0.00897213	0.00450227	0.0108967
The observed distribution	0.0458639	0.0388328	0.0406016	0.0457613

considering both freshness and oldness, we can reproduce a network with a similar distribution.

References

Akoglu L, Mcglohon M et al (2008) RTM: laws and a recursive generator for weighted time-evolving graphs. In: Eighth IEEE international conference on data mining (ICDM), pp 701–706

Alber R, Barabási AL (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Bu T, Towsley D (2002) On distinguishing between internet power law topology generators. INFOCOM 2002, vol 2, pp 638–647. doi:10.1109/INFCOM.2002.1019309

Chakrabarti D, Zhan Y et al (2004) R-MAT: a recursive model for graph mining. In: 4th SIAM international conference on data mining

Du N, Faloutsos C, Wang B, Akoglu L (2009) Large human communication networks: patterns and a utility-driven generator. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '09). ACM, New York

Erdős P, Rényi A (1959) On random graphs. Mathematicae Debrecen 6:290–297

Kleinberg JM, Kumar R, Raghavan P, Rajagopalan S, Tomkins AS (1999) The web as a graph: measurements, models, and methods. In: Proceedings of the 5th annual international conference on computing and combinatorics (COCOON '99)

Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. Science 311(5757):88–90

Leskovec J, Faloutsos C (2007) Scalable modeling of real graphs using Kronecker multiplication. In: Proceedings of the 24th international conference on machine learning (ICML '07). ACM, New York

Leskovec J, Chakrabarti D, Kleinberg J, Faloutsos C (2005a) Realistic, mathematically tractable graph generation and evolution, using Kronecker multiplication. In: European conference on principles and practice of knowledge discovery in databases (ECML/PKDD)

Leskovec J, Kleinberg J, Faloutsos C (2005b) Graphs over time: laws, shrinking diameters and possible explanations. In: ACM SIGKDD international conference on knowledge discovery and data mining (KDD)

Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: densification and shrinking diameters. In: ACM transactions on knowledge discovery from data (ACM TKDD), vol 1, no. 1

Mahadevan P, Krioukov D, Fall K, Vahdat A (2006) Systematic topology analysis and generation using degree correlations. In: Proceedings of the 2006 conference on applications, technologies, architectures, and protocols for computer communications (SIGCOMM '06). ACM, New York

Newman MEJ (2003) Random graphs as models of networks. In: Bornholdt S, Schuster HG (eds) Handbook of graphs and networks. Wiley, Berlin

Pedarsani P, Figueiredo DR, Grossglauser M (2008) Densification arising from sampling fiex graphs. In: ACM SIGMETRICS performance evaluation review—SIGMETRICS '08, vol 36, no. 1

Pennock DM, Flake GW, Lawrence S, Glover EJ, Lee Giles C (2001) Winners don't take all: characterizing the competition for links on the web. Proc Natl Acad Sci USA 99(8):5207–5211

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on world wide web (WWW '10). ACM, New York

Vázquez A (2003) Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations. Phys Rev E 67:056104

Watts DJ, Strogatz S (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442. doi:10.1038/30918