Preferential Attachment in Social Media

The Case of Nico Nico Douga

Johannes Putzke[1] and Hideaki Takeda[2]

[1] University of Bamberg,
An der Weberei 5, 96047 Bamberg, Germany
[2] National Institute of Informatics (NII)
2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan

Abstract. In the examination of evolving complex networks, the analysis of preferential attachment is a core research problem. However, the results of studies that examine whether preferential attachment operates in social media networks are conflicting. On the one hand, preferential attachment generally has been found to be a stable predictor of network evolution. On the other hand, IS researchers question the applicability of the preferential attachment hypothesis to social media networks. This study shows that preferential attachment also operates on Nico Nico Douga, a Japanese video sharing service similar to Youtube with more than 20 million registered users. However, this study also reveals that the attachment kernel differs substantially from the classically assumed log-linear form, when estimating the kernel with nonparametric maximum likelihood estimation (PAFit).

Key words: PAFit, preferential attachment, social media, social network analysis

1 Introduction

Social media radically have changed the way in which we create and consume content. During content creation and consumption we are embedded in networks of peers that leave their digital traces on the social media platforms that we use. Consequently, a number of recent papers applies methods from network science for the examination of social media, e.g. [1]. In this context, authors particularly tried

to answer the research question what might be a good model to describe the evolution of a network. As a general network formation model, preferential attachment (PA) [2] has been found to be a stable predictor for network evolution. PA means that newly arriving nodes in a network connect with a higher probability to those nodes in the network that already have a large number of connections. However, since recent studies by information systems (IS) scholars [1][3] could not find evidence for PA in social media networks, these scholars question the applicability of the PA hypothesis to networks in social media. On the other hand, a recent study about friendship networks on the social media website *Flickr* [4] provides evidence for PA in social media networks, but not in the classically assumed log-linear form of the attachment kernel. Therefore, in this paper we try to shed further light on the question whether PA operates in social media networks. Particularly, we intend to examine whether [4]'s findings can be replicated on a data set from the social media website *Nico Nico Douga (NND)*. *NND* is a Japanese video sharing platform similar to *Youtube*. As of 2014, it has more than 20 million registered users [5]. However, *NND* has a unique feature that differentiates it from *Youtube*. In *NND*, users can add time-stamped comments to the videos. These comments are then overlaid to the original video when playing back the video. This commenting feature made *NND* famous for a new form of content co-creation. In this form of content co-creation, song writers and (3D) illustrates collaborate and create music videos. In doing so, they comment on their videos and (re-) use some content (e.g. music or graphics) from each other. When creating content and using the content of other videos, the creators frequently attribute credits to the videos which they used. In such a way a network of co-creation emerges. We analyze the PA hypothesis for this evolving co-creation network.

The remainder of this paper will be structured as follows. The next section, Background, will be structured into two sub-sections. In the first sub-section, we review the related literature about the PA hypothesis. In the second sub-section, we present [4]'s nonparametric maximum likelihood (ML)-based preferential attachment kernel estimation method (*PAFit*) as well as its application to a *Flickr* dataset. In the next section, Replication I, we describe the *NND* data set used for replicating the results of [4]'s study, the procedures used in the replication, as well as the comparative results of the application of *PAFit* to our data set (see also [6]). Since we could replicate [4]'s results with the *NND* data set, we performed another replica-

tion with a publicly available social media data from the Website *Digg*.[3] This replication will be presented in the section "Replication II". Finally, the paper closes with a short Discussion section.

2 Background

2.1 The Preferential Attachment Hypothesis

The "preferential attachment hypothesis" has been examined in the literature under various names such as the "Yule distribution/process" [7], the "Mathew effect" [8], or "cumulative advantage" [9].[4] It states that subjects with an attribute X will acquire new units of this attribute X according to how many units of this attribute they already have. In network science, "preferential attachment is generally understood as a mechanism where newly arriving nodes have a tendency to connect with already-well connected nodes" [12]. Most researchers attribute the name "preferential attachment hypothesis" to [2] who published a highly influential paper in *Science* about the subject.[5] Considering the vast amount of papers on PA, it would not be meaningful to provide a complete (interdisciplinary) literature review on this subject at this place.[6] Rather, an appropriate literature review focusses on the groundbreaking works about PA (e.g., [13,14]), as well as on the works of IS researchers who conducted network studies and pointed out to the (missing) PA process in the context of social media. Concerning the groundbreaking works about PA, the reader is referred to the literature reviews in [15] and [16], as well as the corresponding sections in the work by [4]. Concerning the works by IS researchers, there were some interesting findings concerning PA in social media. For example, [3] could not find evidence for the PA hypothesis examining data from 28 online communities. Also [1] do not find evidence for PA examining enterprise social media networks such as an online social networking platform. On

[3] `http://digg.com/`, accessed on 02/28/2017.

[4] For the history of the PA hypothesis see also [10] and [11]. The interested reader is also referred to a lecture by Aaron Clauset (available at `http://tuvalu.santafe.edu/~aaronc/courses/5352/fall2013/csci5352_2013_L13.pdf`, accessed on 04/12/2017) in which Clauset explains the PA hypothesis and its history in detail.

[5] The paper has been cited more than 25,000 times as of a google scholar search on 04/17/2017.

[6] For example, a google scholar search for the term "preferential attachment" provided more than 24,000 search results as on 04/17/2017.

the other hand, PA is supposed to be a robust predictor of tie formation [1], and other IS researchers state that, for example, fundraising over social media is a PA process [17]. In the light of these contradicting results, it is evident that we need a clearer understanding about the "conditions under which preferential attachment operates (or not) in different network settings". Therefore, [1] call exactly for this type of research. In order to answer [1]'s call, a robust method for estimating PA in different network settings is needed. Such a method has been recently proposed by [4]. However, this method has never been applied in IS research. Therefore, in the following sub-section we highlight [4]'s method, as well as its application by [4] to a *Flickr* social network dataset [18].

2.2 Nonparametric Maximum Likelihood-Based Preferential Attachment Kernel Estimation and its Application to Flickr

Following [4], we denote an observable seed network at a time-step $t_0 = 0$ with G_0. This network grows from each period $t = 0, 1, ..., T$ with $n(t)$ nodes and $m(t)$ edges. At discrete points in time $t = 0, 1, ..., T$ we can observe these static network configurations G_t. In each time-step t, the probability that an existing node v with in-degree k acquires a new edge is given by

$$Pr(v \text{ acquires a new edge}) \propto A_k. \tag{1}$$

A_k is the value of the attachment kernel at degree k. A number of authors proposed estimation methods for the attachment kernel. A good overview of these methods can be found in Table 1 in [4].[7] However, most of these methods assume a log-linear form $A_k = k^\alpha$ of the attachment kernel. Notable exceptions are the works by [13] and [14] who base their estimation on histograms, and are thus nonparametric. However, also these two methods have their shortcomings. In contrast, [4] derive the ML estimator as[8]

$$A_k = \frac{\sum_{t=1}^{T} m_k(t)}{\sum_{t=1}^{T} \frac{m(t) n_k(t)}{\sum_{j=0}^{K} n_j(t) A_j}} \tag{2}$$

for $k = 1, ..., K$. The solution to this equation can be found using the Minorize-Maximization (MM) algorithm (e.g., [19]) that is beyond the scope of this paper.

[7] doi:10.1371/journal.pone.0137796.t001, accessed on 3/31/2017.
[8] For the details of derivations and proofs of the following paragraph see [4].

The interested reader is referred to the aforementioned literature. [4] apply their method to a publicly available *Flickr* social network dataset [18].[9] This dataset consists of 2,302,925 users and their 33,140,017 directed friendship relationships that grow over a period of 133 days. After the period $t = 0$, 815,867 new nodes and 16,105,211 new edges arrive in the data set. As convergence criterion for the MM algorithm, [4] use a value of $\epsilon = 10^{-7}$. Figure 1 [4] illustrates the results of the estimation of the attachment kernel. The plot is on a log-log scale, and a solid line illustrates $A_k = k$ as a visual guide. Although the results clearly indicate PA, they also indicate a clear signal of deviation from the log-linear model $A_k = k^\alpha$ [4].

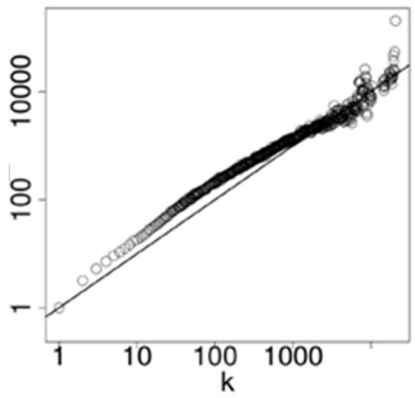

Fig. 1: Estimation of the attachment kernel in the *Flickr* social network dataset (doi:10.1371/journal.pone.0137796.g003)

3 Replication I

3.1 Data Set: Nico Nico Douga

We replicated [4]'s study using a data set of NND that was provided by [5], and that is partially available on figshare[10]. The data set contains the metadata of all videos uploaded on *NND* between January 2007 and December 2012 (i.e. the author, keywords, author's comment, number of views and the timestamp of the upload). In

[9] Available at `http://konect.uni-koblenz.de/networks/flickr-growth`, accessed 02/28/2017.

[10] Available at `https://dx.doi.org/10.6084/m9.figshare.2055597`, accessed 02/28/2017.

total, we extracted 2,622,495 VideoIDs from the data set that had at least one keyword associated to them, together with their timestamps. Out of these 2,6 million videos, 1,427,715 videos could be assigned to an author ID (see [5]). Our following analyses are based on these 1.4 million videos.

3.2 Methods

For the estimation of the *PAFit* model, we focused on the author co-creation network, i.e. we assumed a directed link from author *A* to author *B*, if an author *A* cited a video that had been created by author *B*. In this way, we obtained 4,773,163 directed edges. After excluding self-citations, 3,014,423 edges remained in the data set. When estimating the model in *R v. 3.2.2* with the package *PAFit v. 0.9.3* on the whole data set, we obtained an error due to memory problems. Therefore, we decided to split the data set into two (random) parts. The first part contains 2,016,458 random edges, the second part contains the remaining 997,965 edges. The size of the first part is the maximum number of edges for which the estimation worked on our system. The first part of the data set will be used for model estimation, and the second part of the data set will be used for cross-validation. The final data set consists of 124,996 authors and their relationships that grow over a period of 1,449 days. In the first part of the data set, after the period $t = 0$ 115,134 new nodes, and 1,635,827 new edges arrived. The node with the highest number of edges has a degree centrality of 38,234. In the second part of the data set, after period $t = 0$ 115,134 new nodes, and 808,740 new edges arrive. The node with the highest number of edges has a degree centrality of 19,037. Like [4] we use a value of $\epsilon = 10^{-7}$ as convergence criterion for the MM algorithm. Furthermore, we use logarithmic binning (with 200 bins) in order to stabilize the estimation of the attachment kernel.

3.3 Results

Figure 2 (a/b) illustrates the results of the estimation of the attachment kernel. Again, the plots are on a log-log scale, and solid lines illustrate $A_k = k$.

The estimated attachment exponents of the log-linear model $A_k = k^\alpha$ are (1) $\alpha = 0.8819457$ for the first part of the data set, and (2) $\alpha = 0.8841727$ for the second part of the data set. In summary, these results are very stable, and provide strong empirical evidence for preferential attachment in the *Nico Nico Douga* co-

Fig. 2: **(a/b).**Estimation of the attachment kernel in the *NND* data set (first part, second part)

creation network. Nevertheless, like Pham et al. [4] we observe a deviance from the log-linear functional form of the attachment kernel, particularly in the high degree region.

4 Replication II

Since we could replicate [4]'s results with the *NND* data set, we performed another replication with a publicly available social media data. This data set comprises user interactions on the social media website *Digg* between 10/28/2008 and 11/12/2008.[11] In the data set, each node reflects a user in the network, and an edge between user *A* and user *B* reflects that user *A* replied to user *B*. The data set consists of 30,398 nodes, and 87,627 edges between them. Concerning the evolution of the network, there was only one edge with a timestamp 10/28/2008 in the dataset. Therefore, we assumed that the observable seed network at a time-step $t_0 = 0$ comprises all edges with a time-stamp ≤ 10/29/2008. During the evolution of the network, 30,382 new nodes and 59,655 new edges arrived. Since the maximum degree of the nodes was rather low (243), we did not apply logarithmic binning.

Figure 3 displays the estimation results. The estimated attachment exponent has a value of $\alpha = 0.3958123$. Again, the results indicate PA, but also not in log-linear form. The low value of the attachment exponent α is an interesting finding, particularly since the social news website *Digg* is used for professional as well as for private use.

[11] Available at `http://konect.uni-koblenz.de/networks/munmun_digg_reply`, accessed 02/28/2017.

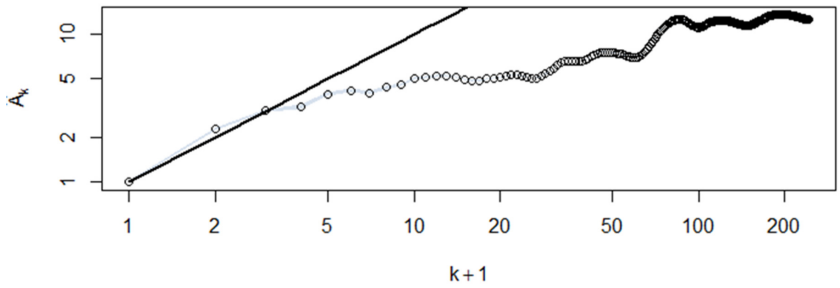

Fig. 3: Estimation of the attachment kernel in the *Digg* data set

5 Discussion

In this paper, we showed that the evolution of the co-creation network on *NND* is driven by PA. However, nonparametric ML-based estimation of the PA kernel revealed that the process does not follow the classically assumed log-linear form. Hence, this study makes at least the following contributions to the IS literature: First, we introduced a new method for attachment kernel estimation, *PAFit* [4], from physics to the IS literature. This is important, as our results show that the predominant praxis to estimate the attachment kernel with parametric methods falls short. Second, IS researchers argue that PA might be a structural feature that operates in a variety of physical and technical networks [20], but question the applicability of the PA hypothesis for social networks in social media (e.g. [1,3]). This study showed that PA also operates in social media networks such as the *NND* co-creation network. This is an interesting finding as it substantiates [1]'s call for research that we should figure out the conditions under which PA operates in social media. The proposed method, *PAFit* can help us to fulfil this aim. Using *PAFit*, we also made an interesting second finding. Although we could observe some deviance from the log-linear model in the *NND* data set, the deviance was even more pronounced in the *Digg* data set (see Figure 3). Despite the large deviance, however, there was a strong evidence for PA. Nevertheless, we suggest that future research should examine the functional forms of the attachment kernel for different social media data sets. In an exploratory study, future research should particularly figure out the conditions when the PA hypothesis holds in social media. For example, based on the results of the *NND* study and the *Flickr* study [4] one might speculate that the PA hypothesis holds in social media settings that focus

on the creation of artistic goods (such as photos and videos), which people mainly use during their free time. On the other hand, based on the analyses of technology related discussion forums [3] and enterprise social media platforms [1] one might speculate that the PA hypothesis does not hold in social media settings that focus on increasing productivity (in enterprises). However, these conjectures still have to be substantiated by examining more social media data sets. We hope that this study will lie the basis for more work into this direction.

Acknowledgments. This work was supported by a fellowship within the FITweltweit programme of the German Academic Exchange Service (DAAD). The authors thank Remy Cazabet for the provision of the *Nico Nico Douga* data set.

References

1. Kim, Y., Kane, G.: Online Tie Formation in Enterprise Social Media. In: ICIS 2015 Proceedings. (2015)
2. Barabási, A.-L., Albert, R.: Emergence of Scaling in Random Networks. Science 286, 509–512 (1999)
3. Johnson, S.L., Faraj, S., Kudaravalli, S.: Emergence of Power Laws in Online Communities: The Role of Social Mechanisms and Preferential Attachment. Mis Quart 38, 795–808 (2014)
4. Pham, T., Sheridan, P., Shimodaira, H.: PAFit: A Statistical Method for Measuring Preferential Attachment in Temporal Complex Networks. Plos One 10, 1–18 (2015)
5. Cazabet, R., Takeda, H.: Understanding massive artistic cooperation: the case of Nico Nico Douga. Social Network Analysis and Mining 6, 1–12 (2016)
6. Niederman, F., March, S.: Reflections on Replications. AIS Transactions on Replication 1, paper 7, pp.1–16 (2015)
7. Simon, H.A.: On a class of skew distribution functions. Biometrika 42, 425–440 (1955)
8. Merton, R.K.: The Matthew effect in science. Science 159, 56–63 (1968)
9. Price, D.d.S.: A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science 27, 292–306 (1976)
10. Newman, M.E.J.: Networks: An Introduction. Oxford University Press Inc., New York, United States (2010)
11. Barabási, A.-L.: Network science. Cambridge University Press (2016)
12. Kunegis, J., Blattner, M., Moser, C.: Preferential attachment in online networks: Measurement and explanations. In: Proceedings of the 5th Annual ACM Web Science Conference, pp. 205–214. (2013)
13. Newman, M.E.J.: Clustering and preferential attachment in growing networks. Phys Rev E 64, 025102-1–025102-4 (2001)

14. Jeong, H., Neda, Z., Barabasi, A.L.: Measuring preferential attachment in evolving networks. Europhys Lett 61, 567–572 (2003)
15. Barabasi, A.L.: Scale-Free Networks: A Decade and Beyond. Science 325, 412–413 (2009)
16. Hidalgo, C.A.: Disconnected, fragmented, or united? a trans-disciplinary review of network science. Applied Network Science 1, 6 (2016)
17. Tan, X., Lu, Y., Tan, Y.: An Examination of Social Comparison Triggered by Higher Donation Visibility over Social Media Platforms. In: ICIS 2016 Proceedings. (2016)
18. Mislove, A., Koppula, H.S., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Growth of the flickr social network. In: Proceedings of the First Workshop on Online Social Networks, pp. 25–30. (2008)
19. Hunter, D.R., Lange, K.: A tutorial on MM algorithms. Am Stat 58, 30–37 (2004)
20. Faraj, S., Kudaravalli, S., Wasko, M.: Leading Collaboration in Online Communities. Mis Quart 39, 393–412 (2015)