Summary Review Documentation for

"Multi-scale Dynamics in a Massive Online Social Network"

Authors: X. Zhao, A. Sala, C. Wilson, X. Wang, S. Gaito, H. Zheng, and B. Y. Zhao

Reviewer #1

Summary: This paper analyzes the dynamics of a really popular social network in China called Renren. It analyzes three aspects of the social network, user and edge formation dynamics, community dynamics, and finally, an interesting event where two different social networks merged together. Several interesting findings including the validity of preferential attachment model during the evolution of the network have been reported.

Strengths: Nice paper with nice data set and nice results. Well written paper.

Weaknesses: Some minor clarifications are required, but nothing major.

Comments to authors: I think it is commendable that the authors went ahead and found such a nice data set. It would be a nice asset to the community if they make it publicly available for research purposes. The fact that the data included a rare merger event was even more interesting since it is usually part of real life, but hard to capture such data.

The initial portion of the paper was very nice with very interesting findings about the evolution of the social network and showed nice graphs that show how things more or less stabilize after a certain critical mass of users are accumulated.

I also liked the section 3.2 that talks about preferential attachment model and its applicability on the Renren data. It also shows some nice properties like after the social network becomes larger, its edge creation is no longer driven solely by preferential attachment model alone. I am not sure of the particular reasoning that somehow it becomes harder to locate supernodes inside the massive network. I don't think preferential attachment necessarily means that users are trying to locate supernodes; I think it could also the opposite where supernodes tend to be looking out for friends and such, since this is a undirected graph as opposed to a directed graph. In any case, I am not sure the reasoning is correct and might be nice to clarify this.

The one section I had the most trouble with is the section 4. First, I am not sure what the definition of a community is in graph theory terminology. Are you talking about strongly (weakly) connected components ? If yes, why not state this explicitly. In any case, a more precise definition of community is required before I can appreciate this section better.

Similarly, it might be illustrative if you can spare a few sentences on describing modularity and other terminology introduced in this paper more precisely. This is important especially for non experts to not have to refer to external material for understanding what you have done.

I think the observation that with 99% probability, a community i will merge with another community that has the largest number of edges to i is kind of directly related to the definition of community. Community typically refers to a set of nodes that have large number of links amongst themselves, so it is only natural to merge with something to which there are a large number of edges to begin with. That is even more the reason why my earlier comment is important.

Reviewer #2

Summary: The paper explores a new large-scale longitudinal dataset of the Renren social network with several hundred million users. A particular focus is on dynamics, including the evolution of community structures and their impact on the edge generation process. Also, a unique merger of two OSNs is captured and explored.

Strengths: The dataset considered here is certainly quite unique, and the findings should be of some interest to the community, even if they are not particularly surprising. The paper is well written and teases out a number of interesting facts about this social network.

Weaknesses: The paper lacks an overall goal and crisp conclusion; rather, it covers a broad range of statistics and observation, only some of which are surprising. The whole Section 4 on community dynamics depends on the behavior of the community detection algorithm employed; we don't know if some of the features (e.g., the increase in community sizes as the network grows) are due to the algorithm, or if they are in fact "real". In the absence of a ground truth, the authors should have been more careful with methodology and conclusions (see details).

As an aside, there is no indication that this dataset will be made available to the research community. Hence, this paper reports irreproducible research, which has recently become a sensitive topic in our community.

Comments to authors: The paper reports a remarkable set of results obtained from a unique dataset. If the dataset could be released along with the findings, it would be a tremendous resource for the community. However, I feel that a synthesis has only partly been achieved.

My greatest methodological complaint about the paper concerns the study of the impact of communities. You use communities that are defined implicitly as whatever the clustering algorithm outputs. It is very unfortunate that no ground truth was available (at the very least a training set over which to train the parameters, such as delta). We therefore don't know if your conclusions are features of the algorithm or of the actual community structure present in the data. For example, the observation in Fig. 5a that community sizes increase with the network size could very well be a feature

of the community detection algorithm, in that it might favor the larger among all the candidate communities (I'm not familiar with the details of the algorithm and am speculating to make my point). The authors could have stress-tested their results, e.g., by running the same algorithm over synthetic data that has controllable community sizes by design. Unfortunately, this issue casts uncertainty over all the conclusions in Section 4.

Section 3.2: It is my understanding that the seminal work by Anderson, Doyle, Willinger et al. has soundly debunked the applicability of the PA process to model real complex networks, such as social networks. While the rate of edge creation by new nodes is of course of independent interest, it might be useful to clarify this.

Page 10: it seems obvious that if one considers larger communities, a larger fraction of interactions happen inside rather than across these communities. Wouldn't the opposite finding be really puzzling?

The study of the network merger in Section 5 is of course quite a unique event. However, the insights are perhaps more of a business than scientific interest.

Reviewer #3

Summary: This paper examines the growth of the RenRen online social network using data provided by the site (thereby eliminating any measurement bias). The paper focuses on growth effects at three levels: individual nodes (e.g., preferential attachment), community-based (e.g., creating links within the local community), and network-wide (e.g., effects due to network-wide events like merging with another OSN). The paper makes the conclusions that new user activity is not the dominant growth mechanism after the network matures, that preferential attachment varies over time, that users in larger communities are more active, and that the other network that was merged was quickly absorbed and integrated.

Strengths: The paper uses real-world data free of measurement bias, and does a good job of exploring the characteristics of network growth. The paper is also very well written, is clear, and is a pleasure to read.

Weaknesses: My primary concerns with this paper revolve around (a) the novelty of the results (i.e., there have been a number of papers in this space, and I'm a little unclear what insights this paper provides beyond existing work), and (b) the conclusions drawn in the community section (i.e., these conclusions seem to this reviewer to be a consequence of how communities are defined).

Comments to authors: I very much enjoyed reading this paper. It uses a unique and very nice data set, performs interesting analysis on the data, and is overall a pleasure to read.

My primary concern with the paper is over the community section (Section 4), where it seems to me that many of the conclusions that you draw are a consequence of the way you define communities. Let me explain. The Louvain algorithm (like all such modularity-maximization algorithms) finds groups of nodes that are more densely connected internally than the surrounding graph. As a result, high-degree nodes are very, very likely to end up in communities (as leaving them outside of a community would increase modularity much more than leaving out a low-degree node).

Now, with this in mind, your conclusions in 4 would seem to be more a consequence of the community detection algorithms' mechanism than of user behavior. For example, the conclusion "the membership to a community has a significant influence on users' activity. [such users] ... create edges more frequently ..." would seem to inappropriately contribute a cause-and-effect relationship to community membership.

Following up on the previous point, for your goal, I feel that using a graph-based notion of community may be inappropriate. What you're trying to do is to measure the influence that "groups" of users (in the sense of offline communities) have on network structure. Basing that definition on network structure itself would seem to result in circular feedback no matter what you do. For example, the Louvain algorithm detects communities in RenRen with over 100,000 nodes! It would be hard to argue that these form a "community" in the sense of the word that you're interested in. Instead, using of user-defined groups (e.g., Facebook-like groups) would seem to get at what you care about more directly, and would avoid the circular feedback of defining communities based on network structure and then analyzing the effect that communities have on network structure.

My secondary concern with this paper is over the contribution beyond prior work. Without a doubt, this paper has a much richer and reliable data set than prior work. But, some of the conclusions are not particularly surprising (as the authors admit), and I have some doubts about the community section (see above). Overall, I found the most interesting part of the paper to be the network merging section.

I wondering if the changing strength of preferential attachment in 3.2 might be due to another effect (in addition to the visibility issue you bring up): the 1,000 link cap. Intuitively, it would seem that this cap would affect very few users initially, but would eventually start to effect more and more users, making them unable to accept new links. As a result, the "preferential attachment" would appear decreased, as fewer links go to these guys. Could you validate this, by examining the fraction of nodes at the 1,000 degree cap over time (you could convince me this effect isn't there if there are few such users)?

I was a bit unclear about what the takeaway point was in 3.1 (Edge creation). Is the difference in exponents statistically significant?

I was somewhat surprised by the variance in Figure 1(c). It appears that there are a number of events that cause new/old users to create more links (e.g., around days 275, 475, and 650, the older users appear to be creating more links). This suggests to me that there are external forces at work (as you mention before, advertising by RenRen, etc), and these forces seem to have a significant effect on where links are created. A more in-depth discussion of these events would likely be enlightening.

In 3.2, you state that the random/higher degree choices represent upper/lower bounds for estimating $p_e(d)$. But, wouldn't choosing "Dest: Lower Degree" be a more accurate lower bound? "Dest: Random" would seem to represent something in the middle.

In 3.3, you conclude that "edge creation at early stages is driven my new node arrivals, but this decreases". I'm wondering if this is instead a consequence of the decreasing relative network growth rate (shown in Figure 1 b). Since the network seems to be growing at a relatively lower rate over time, wouldn't this naturally result in more links (relatively) being created by older nodes, assuming nodes create links at a constant rate?

Have you considered using a local notion of community (e.g., using some local community detection algorithms) rather than a global algorithm? This might also address my points from above.

The results in Section 4.4 would also seem to suffer from the community comments above. For example, the results in Figure 7a would seem to follow from the fact that more active users are more likely to end up in communities. (similar comments hold for 7b and 7c). Similarly, the results at the end of 4.4 which state that users in larger communities have a higher fraction on in-community links would be expected, no? Simply by statical arguments, if node A with degree X is in a 100-node community and node B also with degree X is in a 100,000-node community, we would expect a higher fraction of B's links to be within B's community, relative to A's links within A's community.

I'm curious about how RenRen could tell when a user had a duplicate account in 5Q and Xiaonei – was this based on name matching? Or, would the user indicate that they had 2 accounts after the merge?

Reviewer #4

Summary: This paper studies the dynamics of the Renren social network at user (edge), community, and network levels. At the node (edge) level, it's shown that most edges are created earlier at the node's lifetime, and that the preferential attachment model weakens as time goes. At the community level, it's observed that renren has a clear community structure, which follows a power-law distribution for community size, and that smaller communities merge quickly into larger communities. At the network-level, the authors study the merge of Renren and Q5, and observe quick merge, preferential attachment towards new nodes, and active communities (renren users) influence less active users (Q5 users).

The paper is well written, although in many spots it lacks deep analysis and strong insight beyond reporting measurements. Although using large unsampled social network from its start till its matureness to understand dynamics is interesting in itself, many of the observation made in the paper aren't novel, and are very well known in prior works in either static or dynamic social networks analysis literature. The choice of certain algorithms to use in this study is unexplained, and the results obtained using them are not clear how they would compare to results if the authors change such algorithms (see details below).

Strengths: Uses large unsampled social graph over a relatively long time to study dynamics. Studies dynamics at different levels, and show different observations at each level. Observes an interesting social phenomenon of social network merger and shows several interesting observations associated with it.

Weaknesses: Relies on a single dataset, thus observations are hard to generalize. In particular, the analysis at the so-called "network scale" is based on a single unique event, namely, the merging of two social networks. Several interesting results lack explanation. Several observations lack novelty, as they are previously reported in the literature of either dynamic or static social networks analysis

Comments to authors: This paper studies the dynamics of the Renren social network at user (edge), community, and network levels. The dataset used in this study captures two years of Renren, from its start until the end of 2007, and includes one interesting social phenomena of social networks merger. At the node (edge) level, this work shows that most edges are earlier at the node's lifetime, and that the preferential attachment model weakens as time goes. Both conclusions are intuitive, and the first finding agrees with prior work (e.g., Mislove et al, WOSN 2008). At the community level, the authors observe that renren has a clear community structure, which follows a power-law distribution for community size, and where smaller communities merge quickly into larger communities. At the network-level, the authors study the merge of Renren and Q5, and observe quick merge, preferential attachment towards new nodes, and active communities (renren users) enforce less active users (Q5 users) to be more active.

The paper is well written, although in many spots it lacks deep analysis and strong insight beyond reporting measurements. Although using large unsampled social network from its start till its matureness to understand dynamics is interesting in itself, many of the observation made in the paper aren't novel, and are very well known in prior works in either static or dynamic social networks analysis literature. The choice of certain algorithms to use in this study is unexplained, and the results obtained using them are not clear how they would compare to results if the authors change such algorithms (see details below).

One big concern I have is related to the dataset used in this study. First, it is a single dataset, and observations and conclusions made in this work give the impression that this would apply to social networks in general. However, that is unsupported claim; in fact the authors at some points (e.g., before the end of page 4) are unsure whether the observation concerning edge dynamics would apply to today's renren, i.e., after 4.5 years of the last logged event in the dataset used in the paper.

In section 4, the authors' use a similarity based community tracking, although this is one of many algorithms that can yield different results of communities. The authors, a) don't support their choice of such algorithm (perhaps the obvious reason why it's used is its simplicity and efficiency to compute and track communities over such large dataset), and b) don't discuss how changing such algorithm would affect conclusions made in this section/paper on the community-level dynamics.

In section 4.2, the authors observe that as time goes, large communities dominate the network, smaller communities merge into larger communities, and distinction between communities fades, but they do not seem to provide any explanation for that. This point should be further elaborated, as it's very interesting. On the one hand, this contradicts with many classical social networks community's studies. On the other hand, this can be perhaps explained by tracking what kind of links bridges such communities, and whether these edges are noise or real ties.

Observations made in section 4.4 are not explained (mainly the last conclusion in 4.5, which I find most interesting). In that direction, I have two suggestions. First, the observation is at an aggregate level, so perhaps looking deeper in characterizing users who generate such influence, increased interactions, etc, would be a good way of explaining the dynamics in the influence at the community level. Second, the influence is attributed in its entirety to the community structure, and the authors omit side effects. For example, much of that influence can be attributed to how recommendations are made to users in large and small communities (versus stand-alone users), and the likelihood that such recommendations would result in real dynamics.

One last point: since Renren limits the number of friends of each user to 1000 (thus the degree of the node is bounded by 1000), it is bit odd to talk about "power law" degree distribution, and considering "strength of preferential attachment" in Section 3.2. Or at least, one should consider a "truncated" version of the power law distribution, and likewise, a "truncated" version of the preferential attachment model.

Reviewer #5

Summary: This paper analyzes how a large social network in China, RenRen, evolves over time. The authors examined the social network dynamics at both the node level and the community level. They also had the unique opportunity to study an earlier social network merge event where two individual networks Xiaonei and 5Q merged to become the later social network RenRen. A main finding from this work is that the popular preferential attachment model used to explain social network growth does not fit the growth of the RenRen social network. Another finding is that the evolution of the community structure in the RenRen social network is highly predictable.

Strengths: This work uncovered a few social network dynamics not known before. The paper is well written. The authors did a nice job in analyzing the various aspects of the network dynamics in detail.

Weaknesses: It is too broad. Each individual topic is not treated in depth so it is difficult to generalize from the results.

Comments to authors: Your work touched upon a few very interesting topics, for instance, how the community structure evolves over time in a large social network. You also had the unique opportunity to study how communities merge after two independent social networks merged. However, I feel you moved too quickly between topics. The paper can be made much stronger if you focus on just one aspect of the dynamics, say the dynamics of the community structure, or even the node-level dynamics, and its implications to graph modeling.

For the first topic, you may compare the results from different community detection algorithms, and use them to justify the choice of the incremental Louvian method you chose. For the second topic, you may propose a new model that can better explain how a social network grows and later work can use your model to generate synthetic social graphs. If you dive deeper in either of these topics, I think either of them could make a solid submission, for the contribution of the dynamic community detection algorithm or the graph model. Presently, although it's interesting to read about the results, it's difficult to distill the general lessons.

In Section 4 it'd be useful to define what you mean by a community. Your results showed that most communities have a very short life time. I find this result counter intuitive and could be an artifact of your community detection algorithm. User communities in the real world should be much more stable, as social relationships are relatively stable. It'd be useful to try other community detection algorithms, and see whether you obtain consistent results.

The incremental Louvian method you described appears to be a new community detection algorithm. I would emphasize this contribution, describe it in more detail, and compares it with other existing methods.

In Section 4.3, "Thus, when a community splits into smaller communities..." It'd be useful to explain why this happens, and what real-world phenomenon it corresponds to. I suspect it is an artifact of your community detection method.

In Section 4.4, I find the conclusions in this subsection problematic. OSNs may have a large number of fake accounts. The users outside of any community could either be isolated real users or fake accounts. Therefore, the statistics you obtain on those users may not reflect what impact community has on real users.

Response from the Authors

We thank the reviewers for their thoughtful comments. Here we summarize the comments and give our response.

In Section 2, we look for the events which caused the faster growth of the average degree around Day 275, 475 and 650. We found that these days were the beginning of the new academic semesters in China. As we observed in Figure 1(a), users of this network got back from home and they became active online again at this time. Thus, they created more edges around this period, which resulted in faster average degree growth.

In Section 3, we address the following questions from reviewers. First, in Section 3.1, we clarify that the exponent of the edge creation gap can be used in an edge creation model but it is difficult to evaluate its significant without a direct comparison to other network data. Second, we explain here the reason why we use the random destinations, not the lower degree nodes, as the lower bound of estimation of $p_e(d)$ in Section 3.2. This is because using lower degree nodes as destinations will skew the result against preferential attachment model. Instead, randomly choosing destinations can represent the worst scenario for the preferential attachment, which can provide us with a reasonable lower bound of the estimation. Furthermore, we clarify that the limitation of 1000 friends for nodes is not the reason for the decrease of preferential attachment. Because the number of nodes with degree 1000 is very small, only 0.0001% of the total nodes in the final snapshot, which hardly has impact on our measurement. Finally, we discuss more about the reason why edge creation is not only driven by preferential attachment shown in Figure 3(c). Intuitively, at the beginning of a social network, since few offline friends of a user are in the network, users can easily pay their attention to the popular supernodes. As the network grows, more and more offline friends of a user get online. Users focus on connecting with people who they may know instead of popular supernodes.

In Section 4, we first clarify that communities in our study are groups of well-connected nodes in the network, which means that there are dense edges inside a community while sparse edges between communities. To address the concern about the potential dependence of our results on Louvain algorithm, we run another non-modularity community detection algorithm to verify our observations about community-level impact. All the results from this new algorithm are shown in Figure 8. The results are consistent with the results from Louvain algorithm. This confirms that communities have positive impact on users' activities. Moreover, we add more text to introduce the background about community detection algorithms and modularity. As far as we know, incremental Louvain algorithm is the only algorithm that can efficiently scale

with our network size, we chose it in our measurement. To help readers understand our results well, we elaborate the reasons behind our observations. We discuss the possible reason behind the splitting of communities in Section 4.3. According to Dunbar's number, users only have limited time or energy to maintain a fixed number of friendships. Thus, the splitting of a community happens when the size of the community grows beyond the number that a user can maintain.

In Section 5, we provide more details about how the network identified duplicate accounts. Our original network used the users' email addresses to determine whether a user had a duplicate account in the second network. If a user used one email to register both networks, the user was allowed to choose which profile he or she wanted to keep when the user first logged in the network after the merge.