Author’s response to reviews

Title: Analyzing hidden populations online: topic, emotion, and social network of HIV-related users in the largest Chinese online community

Authors:

Chuchu Liu (chuchuliu1993@foxmail.com)
Xin Lu (xin.lu@flowminder.org;xin_lyu@sina.com)

Version: 1 Date: 13 Sep 2017

Author’s response to reviews:

Dear Editor,

Thank you for giving us the opportunity to revise the manuscript. We appreciate the helpful feedback provided by all reviewers and editors.

We have revised the manuscript according to your suggestions and have responded to the reviewer’s comments in the attached point-by-point letter.

The main changes in the revision are as follows:

(1) We have added a description to introduce the aim of the manuscript at the end of the Background in the Abstract.
(2) We have replaced “nature language processing” with “natural language processing”, replaced “interactive network” with “interaction network”, replaced “hiv” with “HIV”, and replaced “hot keywords” with “popular keywords”, throughout the paper according to the reviewer’s comments.
(3) We have modified all of the figure captions and added necessary information to improve the readability. And Figure 3 has been re-worked for better interpretation of the interaction network.
(4) In Methods, we have added the description of how the topic modelling were applied, and how the number of clusters was determined. Moreover, we have added a clearer description of the sentiment scoring and the sentiment dictionary.
(5) We have modified the Results and the Discussion sections, and added a new section for the Conclusion.
(6) In response to reviewer 1, we have made a Supplementary Information with the popular keywords and their associated probabilities as two tables in the revision.
(7) We have corrected syntax errors and modified some inappropriate expression in the manuscript.

We hope that we have been able to address all the suggestions raised by you and by the reviewers and believe that the reviewing process have helped in improving the paper substantially. For detailed changes in the manuscript and our replies to all the reviewers’ comments, please see “point by point response to comments by reviewers” in the following pages.

We look forward to hearing from you again.

Best regards,

Xin Lu, Chuchu Liu

Point by point response to reviewers’ comments

Reviewer #1:
Thank you for the opportunity to review this manuscript. The authors present an analysis of an online forum for users discussing HIV, motivated by the need to monitor and understand populations that may be invisible to survey based sampling and the assumption that users who retain some semblance of privacy or anonymity are likely to be more open and honest when discussing sensitive issues online. The manuscript is generally well written and its strengths include the novelty of the application domain, and the scope of the dataset (Baidu Tieba has about as many active users as all of Twitter). I found no major weaknesses in the paper apart from some of the figures being unnecessary or hard to read, and what follows are a set of recommendations for improving the manuscript.
Major issues:

1) Methods: Avoid statements like this: "However, there has been no in-depth examination of the characteristics and dynamics of online community structure for hidden populations." There have been analyses undertaken for hidden populations on sites like Reddit, they are just rare and hard to find because researchers don't describe them as hidden populations. It is also not necessary to make the claim - the research presented here is useful and important regardless.

Response: Thanks for the suggestion. We have modified this description and cited relevant references; see subsection “Community mining”, section “Methods”, page 6.

2) Methods: I have no issues with the selection of the community structure algorithm (Infomap) or the topic modelling, but the manuscript would benefit from a clearer description of how these were applied and how any parameters were selected (e.g. how many topics? how was this determined? I note later that it is 150 but why?). Infomap is meant to produce more consistently-sized communities but doesn't always do this, especially where there are disconnected components. You might wish to compare Infomap to the standard Blondel implementation that is common to many network science software packages.

Response: Thanks for the suggestion. We have added the discussion on the reasoning of selecting Infomap as the community detection algorithm. And details on the topic modelling were updated; see subsection “Community mining”, section “Methods”, page 8.

For topic modelling, we adopted the Hierarchical Dirichlet Process (HDP) model to extract possible topics. HDP is an evolution of the Latent Dirichlet Allocation (LDA) model and can automatically determine the optimal number of topics. In many studies, it is proved that the HDP model can achieve good results in topic modeling. See [1][2], etc.

For the selection of community detection algorithm, we have tested other popular community detection algorithms, e.g. the standard Blondel implementation. And we find that Infomap achieved the best result because the interaction network of the HIV group is directed, weighted and connected. In addition, in a comprehensive comparison study, Infomap was found to be one of the most efficient, reliable and accurate community detecting method, see [3][4].


3) Methods: More detail regarding the use of the sentiment scoring is required. Which package (it's the Stanford dictionary-based one, I assume)?

Response: Thanks for the suggestion. Because the text data collected in Baidu Tieba are in Chinese, we used the most popular Chinese sentiment dictionaries, Hownet and NTUSD in sentiment analysis. We have added corresponding description of the sentiment scoring and the sentiment dictionary used in this study; see subsection “Sentiment analysis”, section “Methods”, page 10.

4) Results: All of the figure captions can be written as full sentences (no need to write "Note:" on a new line), and need to provide enough information to understand how to read/interpret what is in the figure. In addition, Figure 3 needs to be re-worked because they currently provide no useful quantitative information.

Response: Thanks for the suggestion. We have modified all of the figure captions and added necessary information for the stand-alone reading of the figure. Moreover, Figure 3 has been re-worked for better interpretation of the interaction network.

5) Results/Discussion: It would be better if the results section only provided the answers to the questions/experiments described in the methods section, and avoided discussion elements such as comparison to existing literature. The Discussion could be used for that purpose - one paragraph summarising the major findings, one or more paragraphs comparing to existing literature, one or two paragraphs on implications, one paragraph on limitations, and then the conclusion.
Response: Thanks for the suggestion. We have modified the Results and Discussion sections, and added a new section for the Conclusion. In the Results section, we have removed the discussion of the results. And in Discussion section, we have reorganized the discussion as several subsections such that they are divided into mainly three bodies: Summary of main results, Discussion of the main results, Strengths and limitations. See section “Results”, “Discussion” and “Conclusion”.

Minor Issues:

1) Abstract: "Through nature language processing" should be "Through natural language processing"

Response: Text corrected.

2) Background: I suggest starting a new paragraph before introducing the aim of the manuscript at the end of the Background. There are other examples where paragraphs are a bit long.

Response: We have modified the abstract with text indicating the aim of the study. See “Abstract”, page 1.

3) Methods: "interactive network" should be an "interaction network"

Response: Text corrected.

4) Throughout: I think "hiv" should be "HIV" in all instances, even if the website uses lowercase.

Response: Thanks for the suggestion. We have changed the term throughout the paper.
5) Methods: In particular, using two kinds of clustering to detect clusters of topics and clusters of interacting users is a clever idea. You might wish to have a look at one of the papers from my group that does something similar, for inspiration: 10.2196/jmir.6045

Response: Thanks for the suggestion. We have read the suggested paper and quoted it in the corresponding description; see subsection “Community mining”, section “Methods”, page 7.

6) Methods: Replace the use of the term "hot keywords" with a more precise description. Hot is for temperature, not for information entropy or importance.

Response: Thanks for the suggestion. We have replaced all the terms “hot keywords” with “popular keywords”.

7) Results: Figure 6 and Figure 8b do not provide much additional information for the user. I know it can be especially hard to visualise or report word distributions, but I would recommend a supplementary data file with the words and their associated probabilities in column format.

Response: Thanks for your suggestion. We have created a Supplementary Information with the popular keywords and their associated probabilities as two tables in the revision. But we think the wordcloud is helpful for visualizing keywords, and readers can directly and quickly understand which words commonly posted in HIV communities, so we recommend keeping Figure 6 and Figure 8b. However, if both the reviewer and editors prefer to move them into the Supplementary Information, we would be more than happy to do that.

8) Results: Give exact p-values.

Response: Thanks for the suggestion. The p-values presented in the manuscript are generally too small to write their precise digits, for example, the significance for the correlation coefficient (r=0.70) between the average topic similarity and the network efficiency in the text-based
clusters is $p=1.30e^{-23}$. We therefore adopt the common practice of statistic writing for $p$ with three digits, i.e., “$p<0.001$”. And we unified the decimal places of $p$ in the revision.

Reviewer #2:

The study is well motivated. Using data collected from online community to study hidden HIV population is an important topic.

1) The paper is easy to follow, but it will benefit from some professional editing. Some of the language and word choices feel weird. There are quite a few grammar errors as well.

Response: Thanks for the suggestion. We have checked and modified the grammar in the revision, and corrected some inappropriate words, e.g. “hot keywords”, “nature language processing”.

In addition, we have used professional proof reading service from Proof-Reading-Service.com, which is based in England, south of the university city of Cambridge.

2) Missing citations: e.g., "Currently, Baidu Tieba has more than 20 million bars and the number of active users has reached 300 million." Please cite the source of this information. Note this is just an example. There are other places.

Response: Thanks for the suggestion. We have acknowledged and cited the corresponding references in the revision.

3) "all records until August 2016", what is the start time? Since 2003?
Response: The start time is January 2005. We have added this information; see the second paragraph, subsection “Data sources”, page 5.

4) Have the results of the sentiment analysis methods been validated? Sentiment analysis based on word counts is known for its inaccuracy.

Response: Thanks for the suggestion. The rule-based method is often used in sentiment analysis with Chinese text. As we know, the results of this sentiment analysis method are greatly influenced by the quality of the sentiment dictionary. In our study, we used the most popular two Chinese emotional dictionaries, Hownet and NTUSD, whose effects have been well verified in many studies, e.g. “Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon” [1] and “A study on sentiment computing and classification of sina weibo with word2vec” [2].


5) How are the sentiment words extracted? Based on a known sentiment dictionary? Or the authors have reviewed all the words in the entire corpus to assign a sentiment score to every word? It is not very clear in the paper.

Response: Thanks for the suggestion. Because the text data collected in Baidu Tieba are in Chinese, we used the most popular Chinese sentiment dictionaries, Hownet and NTUSD in sentiment analysis. We have added corresponding description of the sentiment scoring and the sentiment dictionary used in this study; see subsection “Sentiment analysis”, section “Methods”, page 10.

6) How is the number of clusters determined? The same case for how is the number of topics determined in LDA?
Response: Thanks for the comments. Since the K-means algorithm needs to determine the number of clusters manually, we use the sum of distances from all nodes to their cluster centers, as a criterion to select the best number of clusters. And we try different number of clusters from 1 to 1000, to find the best model. We have added the description of how the number of clusters was determined in the revision; see subsection “Community mining”, section “Methods”, page 7.

For topic modelling, we adopted the Hierarchical Dirichlet Process (HDP) model to extract possible topics. HDP is an evolution of the Latent Dirichlet Allocation (LDA) model and can automatically determine the optimal number of topics. In many studies, it is proved that the HDP model can achieve good results in topic extraction, see [3][4]. We have modified the description of how the topic modelling were applied and cited corresponding references about HDP; see subsection “Community mining”, section “Methods”, page 8.
