ITCS 6050/8050 PSYCH 6099 Topics in Intelligent Systems: Computational Human Behavior Modeling

Final Project Report **Due: May 3rd, 2018 11:59 pm**

Project Name: Zhihui Liu
Team Members: (if any)

- Introduction (10-15 sentences)
  - Why your topic is important (this should be convincing)
  - Where is it used? Applications
  - Overview of the rest of your paper (section 2 covers...section 3 presents...)

Wikipedia is one of the world's biggest information sources, disseminating information on virtually every topic in the world. The popularity translates Wikipedia to a growing number of articles sources. However, there is the number of vandal users who carry out acts of vandalism on Wikipedia community. The reputation of articles and users of Wikipedia is damaged by vandalism. Digging in the information of Wikipedia data will help people to better understand the human edit behavior on Wikipedia pages and detect vandal editing pattern in advance.

The remainder of this paper is as follows. First, I give a brief overview of background in section 2. Next, in section 3, I present the details of solution for the problem. In section 4, a summary is given. Finally, I formulate conclusions and outline the future work in section 5.

- Background (10-20 sentences)
  - What other people had to say on this topic(s) (be sure to cite your references, and quote as appropriate)
  - You are expected to discuss the books and papers that you include in your references. You must also cite them. If nothing else, include a brief rationale explaining why you thought it was useful.
  - What other people did on this topic (or related topics)
  - Problems and shortcomings of their work
  - How your work is different and better

Wikipedia has been subject to a statistical analysis in some research studies. Dave et al. [5] analyze the history of Wikipedia articles by a visualization tool. Kittur et al. [6] investigate the using of reverting to detect vandalism. But only looking for reverts signaling vandalism is not strict enough to find out most of the vandalism in the history of articles. The tools are using to detecting vandalism on Wikipedia are ClueBot NG and STiki. ClueBot [1] NG uses an artificial neural network to score edits and reverts the worst-scoring edits which is considered as the state-of-the-art bot being used in Wikipedia to fight vandalism. STiki [2] is another tool to help users to revert vandalism edits using edit metadata, user reputation score and textual features. It also leverages the spatio-temporal properties of edit metadata to assign

scores to each edit. Another research [3] use feature from linguistic features and machine learning on a set of 32K edits annotated by humans on Amazon Mechanical Turk. Kumar et al. [4] develop a system, VEWS which based on edit-pairs and edit patterns to study the behavior of vandals on Wikipedia. The behavior of vandals can be identified with those behaviors from benign users. The results show that the combination of linguistic (from ClueBot NG and STiki) and non-linguistic features (from VEWS algorithm) has the best classification performance.
In the vandal data, there are some pages which edited by different vandal users continuously. We can define this action as "group attacking pattern". One difference of my work and previous studies is that I will find out the group attacking pattern by vandal users from the Wiki data set and visualize those patterns. In addition, I will also explore pages have the most edited times in vandal and benign users. The natural language processing method is used to compare the similarity of different pagetitles. This potential information behind data will reveal the taste of users when they are looking for pages to edit.

- Project (at least 25 sentences, feel free to write more)
    - Your approach to the problem
        - What you did, which team member contributed to which part of the problem (make a table here to be concise), include any relevant and specific info. e.g. software packages and data used
    - What did/didn't work?
    - Include graphs, equations, pictures, etc. as appropriate
    - Results (Very important!!)
        - Include relevant observations, measurements, and statistics.

In this project, I use UMDWikipedia data set to explore the human editing behavior in Wikimedia community. The data set consists of all 17,027 vandal users that registered and were blocked by Wikipedia administrators for vandalism between January 01, 2013 and July 31, 2014. There are also a randomly selected of 16,549 benign users who registered between January 01, 2013 and July 31, 2014 in the data set.
For the part of detecting group attacking pattern, I choose the data from January 01, 2013 to December 31, 2013. Table 1 shows data set consists of 11,197 vandal users and 9973 benign users, 42,774 pages that edited by vandal users and 12,584 pages that edited by benign users. Data are stored in csv file. Figure 1 shows the distribution of users and pages. The distribution of vandal and benign users in 2013 is shown in Figure 2. The distribution of vandal and benign pages in 2013 is shown in Figure 3. In the vandal data, there are some pages which edited by different vandal users continuously. We can define this action as "group attacking pattern", one task of this project is to visualize those patterns. In this task, I define ten types of group attacking pattern. The number of groups is listed in the Table 2.

Table 1 General information of the data set

Month	1	2	3	4	5	6	7	8	9	10	11	12	sum
VU	1109	1082	1027	925	1088	721	464	703	874	1066	1243	895	11197
VP	4575	3708	4517	3887	3434	3009	2161	2955	3410	3898	3872	3348	42774
BU	842	789	908	781	756	658	789	830	976	984	844	816	9973
BP	12977	10726	13504	12510	10056	9494	10234	10298	10012	7289	10299	8441	125840

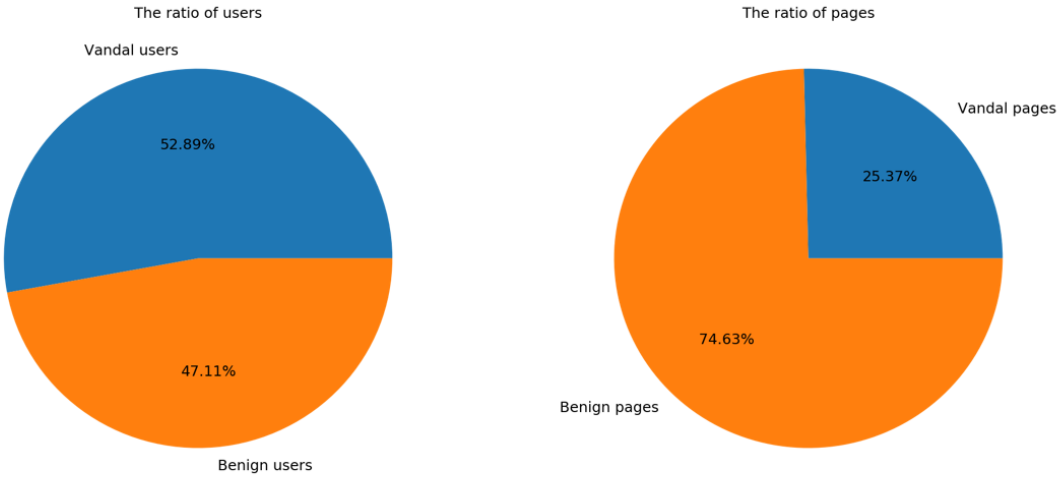

Figure 1 The distribution of users and pages

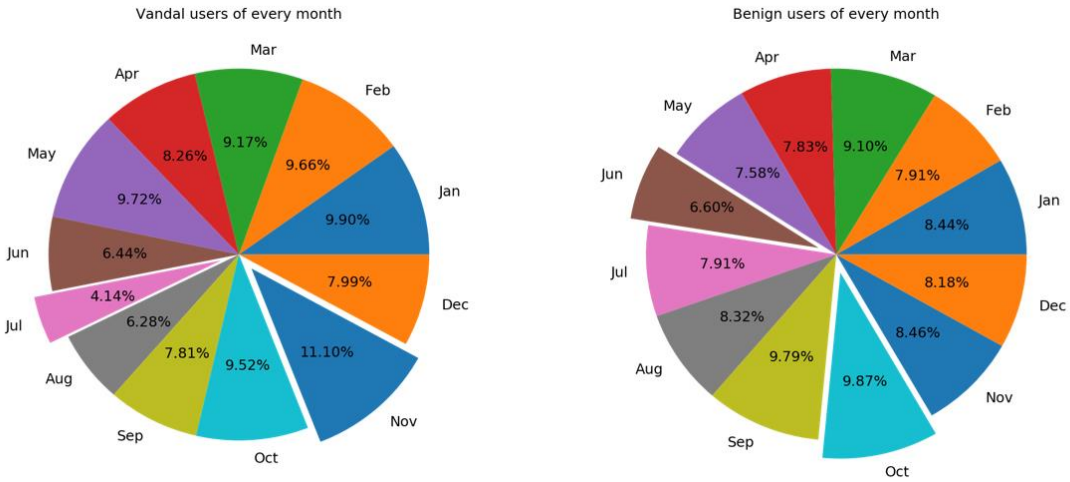

Figure 2 The distribution of vandal and benign users in 2013

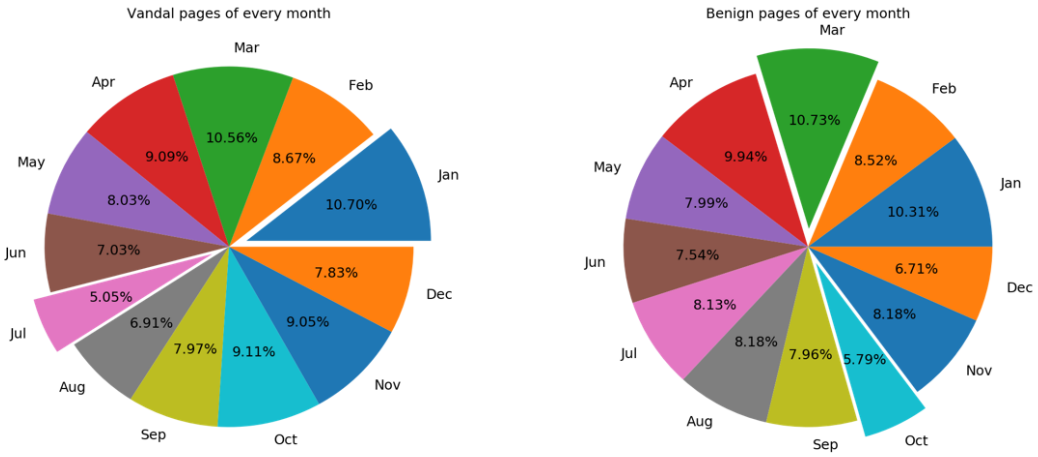

Figure 3 The distribution of vandal and benign pages in 2013

Table 2 The number of different type of group attacking pattern

Type of group attacking pattern	Group number
2 vandal users	176
3 vandal users	19
4 vandal users	17
5 vandal users	1 in Feb, 1 in Nov
7 vandal users	1 in Mar
11 vandal users	1 in Nov
14 vandal users	1 in Nov
18 vandal users	1 in Jan
20 vandal users	1 in Aug
21 vandal users	1 in Aug

The result shows that in all group attacking patterns, there are three main group patterns: 2 vandal users (176 groups), 3 vandal users (19 groups) and 4 vandal users (17 groups). Figure 4 shows the distribution of three main group patterns in 2013. To dig the result of this task deeper, I pay attention to the page titles that are edited by 20 and 21 vandal users. Both of the group patterns are found in August, 2013. And the page titles indicate that there are two events happened in that month. This found can be used to explain that the vandal users editing behavior will be affected by the hot topic in the social media.

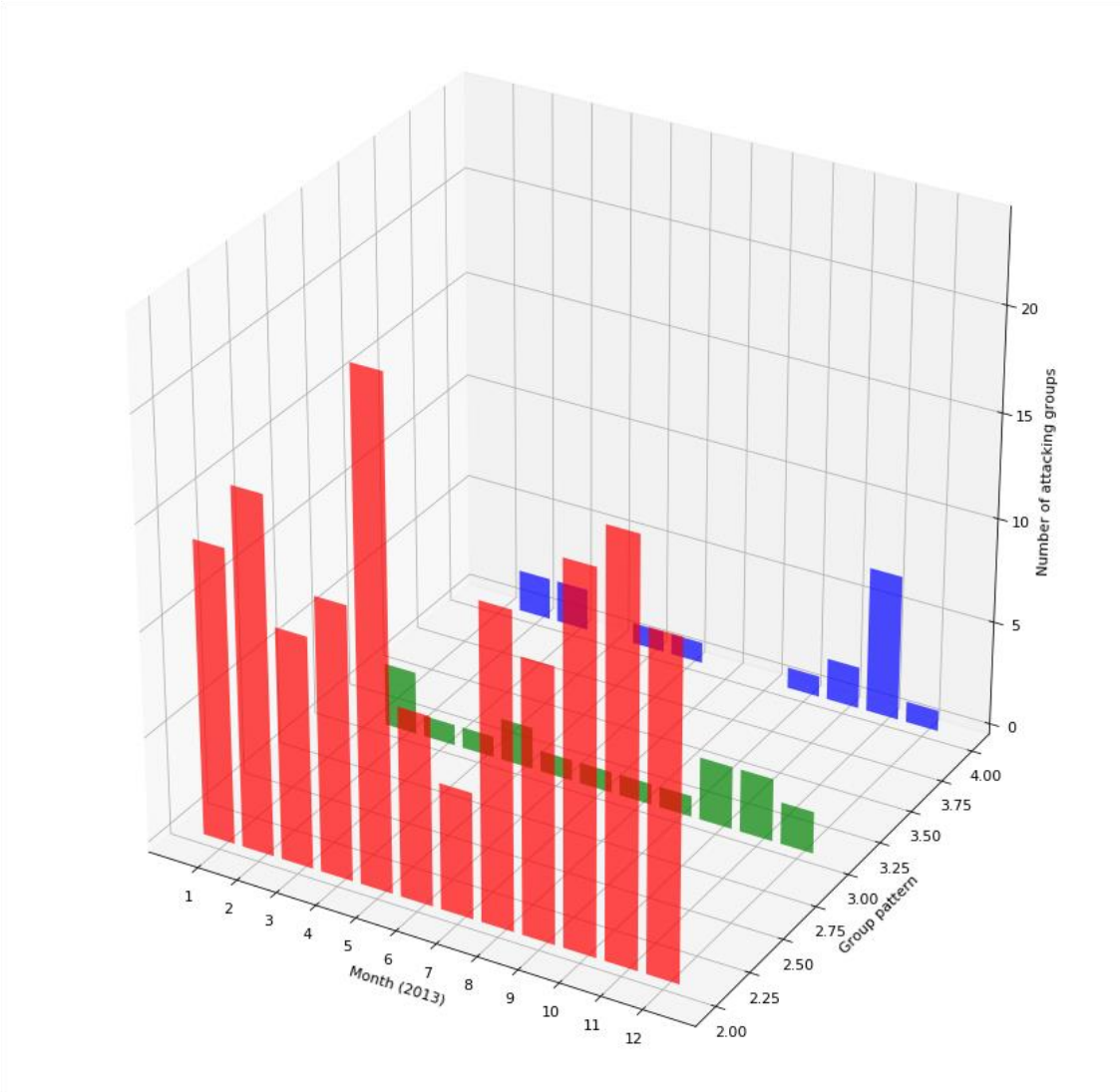

Figure 4 Three group attacking patterns

In the other part of my project, I explore the similarity of semantic in page titles. This work can reflect the different of editing behaviors between vandal and benign users. The potential information behind data will reveal the taste of users when they are looking for pages to edit.

The top 20 edited page titles from vandal and benign users are calculated and one key word that can present the category of page title is extracted. The NLP package, genism is used. I also use word2Vec model to compute the similarity coefficient between each key word. The Google's pre-trained model includes word vectors for a vocabulary of 3 million words and phrases that trained on roughly 100 billion words from a Google News dataset. In this part, data are collected from Jan 1, 2013 to July 31, 2014 which contain 146,522 times edit by vandal users and 610,422 time edit by benign users. Figure 5 lists the top 20 pages that are edited by vandal and benign users. Figure 6 shows the similarity of semantic matrix for vandal editing. Figure 7 shows the similarity of semantic matrix for benign editing.

Algorithm: Detecting Top N page titles

```
1     Detect Top ()
2     {
3         Set count is 1
4         FOR each page title
5             IF Current page title = = Next page title
6                 count ADD one
7             ELSE
8                 IF count greater than N
9                     PRINT OUT “name of page title” and count
10                Set count is 1
11        RETURN Detect Top
12    }
```

Figure 5 Algorithm for detecting top N page title

P1	Bad Girls Club (season 1)
P2	Bad Girls Club (season 2)
P3	Bad Girls Club (season 4)
P4	Bad Girls Club (season 5)
P5	Bad Girls Club (season 6)
P6	Bad Girls Club (season 7)
P7	Bad Girls Club (season 8)
P8	Bad Girls Club (season 9)
P9	Bad Girls Club (season 10)
P10	Bad Girls Club (season 11)
P11	Iron Man in film
P12	Khanpur Ahir
P13	List of Dani's Castle episodes
P14	Nanyang Polytechnic
P15	User talk:Nishidani
P16	User talk:PuffCJMurder123
P17	User:Survivorfan1995/sandbox
P18	Wikipedia:Administrator intervention against vandalism
P19	Wikipedia:Administrators' noticeboard/Incident
P20	Wikipedia:Sandbox

P1	Abdul Khaliq (athlete)
P2	Death of Adrian Donohoe
P3	Draft:Kosi (spiritual teacher)
P4	Gellerup
P5	List of Twenty20 International records
P6	Maciek Pysz
P7	Stephen Jerzak
P8	Talk:List of metro systems
P9	Talk:Star Trek Into Darkness
P10	The Yale Record
P11	User:Maria roswitha wagner
P12	USS PC-598
P13	Wikipedia:Sandbox
P14	User:Bensonfood
P15	Wikipedia:Teahouse/Questions
P16	Wikipedia:Help desk
P17	Wikipedia:Administrator intervention against vandalism
P18	User:Mevank Deonie/sandbox
P19	User:Dr.Gulliver/sandbox/Alexander Selkirk
P20	User:AJGI90

Figure 6 Top 20 pages edited by vandal (left) and benign (right) users

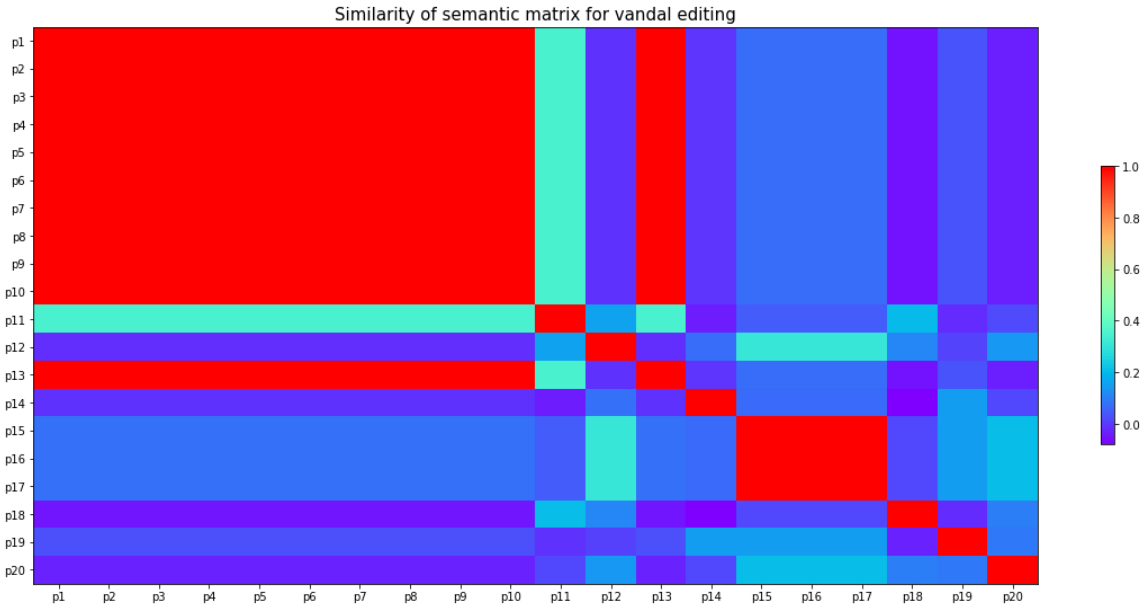

Figure 7 Similarity of semantic matrix for vandal editing

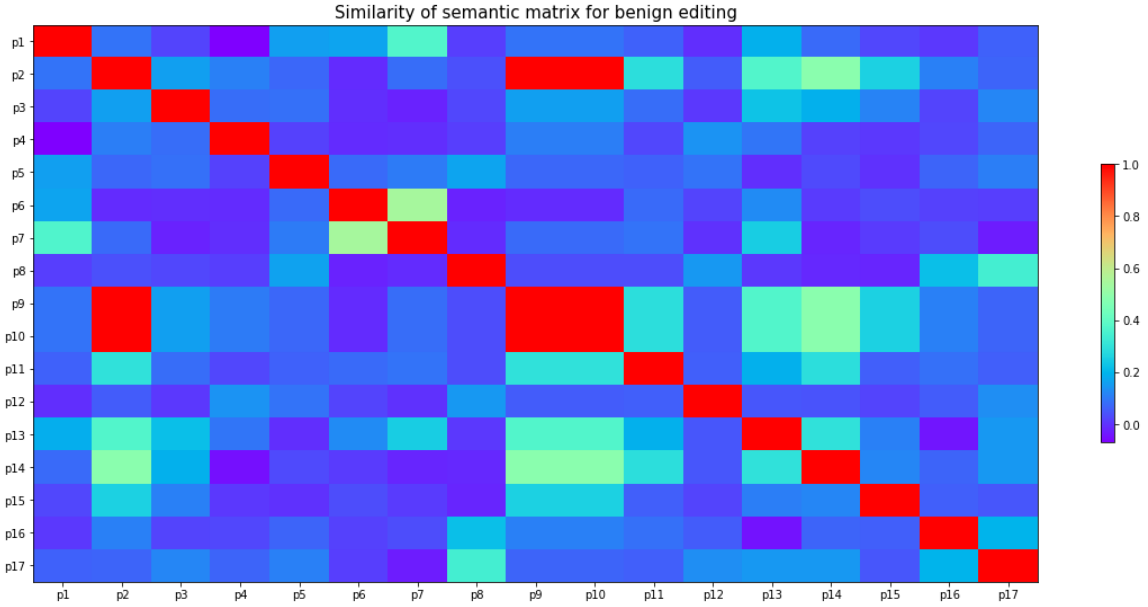

Figure 8 Similarity of semantic matrix for benign editing

From the results of similarity of semantic matrix, we can find that the area of high similarity coefficient is larger in vandal editing than benign editing which shows that vandal users have a more common editing behavior than benign users. By contrast, benign users show a more random editing behavior in the Wikipedia community.

- Summary (10-20 sentences)
    - Try to draw together the intro, background, and project sections.

    - How do they all relate together? (They may appear to be disjoint sections to an unfamiliar reader).
    - Restate important results

Since the inception of Wikipedia in 2001, the free encyclopedia has grown rapidly to become one of the biggest sources of information on the internet. However, there is the number of vandal users who carry out acts of vandalism on Wikipedia community. The reputation of articles and users of Wikipedia is damaged by vandalism.
There are several studies have worked on detecting vandalism in Wikipedia. Kittur et al. [6] investigate the using of reverting to detect vandalism. But only looking for reverts signaling vandalism is not strict enough to find out most of the vandalism in the history of articles.
In my project, one task is to detect the "group attacking pattern" and visualize those patterns. To reveal the taste of users when they are looking for pages to edit, similarity of semantic matrix is created to show the different editing behavior of vandal and benign users. The results show that the vandal users editing behavior will be affected by the hot topic in the social media and vandal users have more common editing behavior than benign users. While, benign users show a more random editing behavior in the Wikipedia community.

- Conclusions (10-20 sentences)
    - What was accomplished / learned
    - What you would have done differently
    - Future work

In this project, I explore the human editing behavior in Wikimedia community. Different from previous works with Wikipedia vandalism problem. I detect the "group attacking pattern" and leverage similarity of semantic matrix to show the different editing behavior of vandal and benign users. The results demonstrate that there are group attacking patterns in the vandalism and these patterns can reflect the taste of vandal users when they are looking for pages to edit. What's more, benign users show a more random editing behavior when they choose pages. Compare to benign users, it seems vandal users have a more common editing behavior. Therefore, digging in the information of Wikipedia data will help people to better understand the human edit behavior on Wikipedia pages and detect vandal editing pattern in advance.

When computing the similarity of key words, a pre-trained model that trained on words from a Google News dataset is used. For future work, I will train my own model that using words from Wikipedia community. Detecting fake accounts from vandal users is another interesting direction to explore.

- References (no limit)
    - You should include papers and articles that were useful. Wikipedia is not appropriate.
    - Cite the papers/books that you used
    - Anything you found useful

[1] http://en.wikipedia.org/wiki/User:ClueBot_NG.
[2] http://en.wikipedia.org/wiki/Wikipedia:STiki.
[3] S.M. Mola-Velasco, "Wikipedia vandalism detection through machine learning: Feature review and new proposals - lab report for pan at clef 2010." in *CLEF*, 2010.
[4] Kumar S, Spezzano F, Subrahmanian VS. Vews: A wikipedia vandal early warning system. InProceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining 2015 Aug 10 (pp. 607-616). ACM.
[5] Viegas, F. B.; Wattenberg, M.; and Dave, K. 2004. Studying Cooperation and Conflict between Authors with history flow Visualizations. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.
[6] Kittur, A.; Suh, B.; Pendleton, B. A.; and Chi, E. H. 2007. He Says, She Says: Conflict and Coordination in Wikipedia. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.