Social network analysis of WallStreetBets Subreddit

**Group 12**

Github repository: <https://github.com/MartinSchmauch/network-analysis-for-reddit>

Martin Schmauch   
University of OuluOulu, Finland  
Martin.Schmauch@student.oulu.fi

Arda Tekin  
University of OuluOulu, Finland  
Arda.Tekin@student.oulu.fi

Nikolaos Georgiadis  
University of OuluOulu, Finland  
Nikolaos.Georgiadis@student.oulu.fi

*Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *\*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*. (*Abstract*)

Keywords—component, formatting, style, styling, insert (key words)

# Introduction (*Heading 1*)

This template, modified in MS Word 2007 and saved as a “Word 97-2003 Document” for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

# Problem Description

## Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the Microsoft Word, Letter file.

## Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

# Dataset Description

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.

Identify applicable funding agency here. If none, delete this text box.

* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# General Methodology

The general procedure for our project starts at looking for suitable posts in the “WallStreetBets” subreddit. We decided to choose the most active subreddits for the last week regarding user interaction in form of comments. Then, we collected the comment data from the webpages related to the posts. From this data we created a network by connecting user nodes by their commenting behavior. Next, we examined the nature of our network and analyzed it using common metrics like diameter, in- and out degree, et cetera. Finally, we summarized our results and drew conclusions based on our observed data.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

**Project 12 Report: Structure**

* **Before Introduction:** Group Leader
* **Introduction:** Introduce your subject and show evidence that you have surveyed any related literature to provide a sound motivation ground and highlight what will be the novelty as compared to other work you surveyed.
* https://medium.datadriveninvestor.com/analyzing-the-r-wallstreetbets-hivemind-84529946f6e8
* **Problem Description:** as you can understand it from the project specification provided to you

How to find the most influential person on the blog/person with most upvotes

* **Dataset Description:**

Dataset scraped using PRAW library that utilizes reddit API. We get comments that contain certain attributes (author, ID, replier, content,...)

* **General Methodology:** outlines the overall methodology and steps taken to solve the problem

Architecture of our project and analysis approach

* **Detailed Methodology:**

Detailed Architecture with some diagrams

* **Results and discussions:** Use a combination of table, curve or both to highlight your findings.. and discuss your results

Show pictures from our graphs, metrics

* **Conclusion and Perspectives**

What did we find out? How can we use this for future work?

* **References**

sources

* **Appendix**

Other content that is not in the overall report

START TEXTS ON NEXT PAGE!

**What we talked about:**

Introduction: Introduction to [Reddit](https://www.reddit.com) and [WallStreetBets](https://www.reddit.com/r/wallstreetbets/), what the other project gone through explain it(related work) [2](https://medium.datadriveninvestor.com/analyzing-the-r-wallstreetbets-hivemind-84529946f6e8)-[3](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4006283) [sources](https://ieeexplore.ieee.org/abstract/document/9540168?casa_token=QyqSjm4fZpEAAAAA:Fh8fSCQX0VD4vLPxHySLc97blHKGXnBecvLhG9GLmtxOKAybq8tLT3BWnaDAEt-xaGzzLGRSVw). Tell what other parts consists of.

Problem description: Find most influential user.

Dataset Description: How we got the data, data schema/types -> preprocessing and graph.gml

General Methodology: What are our general steps.

Detailed Methodology: Formulas, algorithms and descriptions about how we solved the tasks.

Results and Discussions: Describe result tables / graph.

Conclusion:

**Introduction**

Reddit is a website that users can post media, links, photos, polls, and video contents for other people to see. Reddit is a website that can be identified as Forum as well as a blog. Use of reddit is dependent on the user. Some people use the website to talk about themselves while other may use the website to talk about certain topics. The place that these talks happening are subreddits. These subreddits can be identified as small blogs that are within the website for different subject or interests. Users can vote or comment in the posts that are in these subreddits. Vote system helps users to identify good posts from bad ones while we can also use comments to see reaction and interaction of the users. The most upvoted posts considered as top posts in those subreddits. While number of comments may warry from post to post though the top posts usually have the highest number of comments. We wanted to focus on an active subreddit so we can get fresh data from the users. As of 2021, Reddit has more than 2.8 million subreddits and 130,000 active communities. So, we choose the most active subreddit at the time (maybe the time?) which was r/WallStreetBets. WallStreetBets is a subreddit for users to talk about the market and trading. Each day, there are around 800 posts and 50,000 comments (these number may change from time to time). Reddit users can also buy PowerUps for their subreddit of choice these PowerUps makes users comment standout and boosts the communities for 5$ per month or users may prefer Reddit Premium for 6$ per month which gives them few other benefits which include Coins. Coins may be used to gift awards; the award makes the posts standout more or just to reward the user and the community for making a good post.

I will write more here.

**Problem Description**

We needed to create network graph for most active blogs in our case subreddits. Our goal was to find the most influential user in the WallStreetBets. We did not have a ready dataset, so we had to get our own data. So, we get the top posts which are most the upvoted ones. We had two ways to find the most influential user. We could find it by getting upvote counts and the user which most upvotes can be the most influential user or we can look at the comments of the users in top posts and we can decide the most influential user by their interactions with the community. We decided to get the most influential users by their interactions.

**Dataset Description**

For this project, we collected our own data from Reddit using the PRAW (Python Reddit API Wrapper) library. It allows us to access the Reddit API by creating and using a reddit object in Python. We created a submission object with an URL pointing to a reddit page of our desire. This object can be used to extract data from the specific reddit page. Since we are only interested in the relation of the comments on each page, we worked on the submission.comments attribute that provides a list of top level comments (CommentForest). These top-level comments then again contain comment replies (CommentForest) in a forest like structure of trees.

To work with these comments, some preprocessing needs to be done. Some comments of lower levels are mostly hidden to hide complexity of the comment reply feed. On the reddit webpage there is a button stating, “N more replies”, which is represented by a MoreComments object in PRAW. Here we needed to unwrap these comments by showing all comments until the lowest level instead of the MoreComments object. We used the replace\_more method to achieve this for every top-level comment. This resulted in a list of forests with all top-level comments. The comment object had a lot of attributes from which we have listed the most relevant ones in the following.

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| author | Provides an instance of “Redditor”. “Redditor” contains name of the user. |
| created\_utc | Time the comment was created, represented in Unix Time. |
| is\_submitter | Whether the comment author is also the author of the submission. |
| replies | Provides an instance of “CommentForest”. |
| submission | Provides an instance of “Submission”, that the comment belongs to. |

From this list of top level CommentForests we created an edgeList, representing comment replies. Everytime an user replied to a comment, an edge is created containing the name of the users as nodes and their comment reply as directed edge. We also included meta data like the timestamp of the comment creation. The schema of the edgeList looks as following:

[ [ author (String), replier (String), timestamp (String)] ]

This edgeList is used to create the graph and analyze the relations of the users replying to each others comments. Before adding the edges to the graph, we checked, whether a user has replied more than once to a specific other users comments. If that is the case, the weight of this edge will be increased according to the number of replies. The data is stored as .gml file in graph format already. Therefore, we created a graph object and add the edges.

**General Methodology**

The general procedure for our project starts at looking for suitable posts in the “WallStreetBets” subreddit. We decided to choose the most active subreddits for the last week regarding user interaction in form of comments. Then, we collected the comment data from the webpages related to the posts. From this data we created a network by connecting user nodes by their commenting behavior. Next, we examined the nature of our network and analyzed it using common metrics like diameter, in- and out degree, et cetera. Finally, we summarized our results and drew conclusions based on our observed data.

**Detailed Methodology**

How did we create the edge list from the CommenForest?

CommentForest – get replies of each reply author recursively from trees of Comment objects

How did we export the edge list to the graph.gml file

How did we check if a user replied more than once to a specific other users comments

How did we create edge weight

How did we do network analysis in detail?

Read in graph.gml as graph using G = nx.read\_gml(“path\_to\_file/graph.gml”)

Method to extract usernames with highes in-/out- degree

**Sources**

Data Description: how to extract comments: <https://praw.readthedocs.io/en/stable/tutorials/comments.html>

Comment class: <https://praw.readthedocs.io/en/stable/code_overview/models/comment.html#praw.models.Comment>