Social Media Influence Analyzer

Mohammed Z. Guniem  
Department of Computer and Electrical Technology  
*University of Stavanger*Stavanger, Norway  
mghunime@yahoo.no

*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)

Motivation

Upon the rise of the digital revolution through the last two decades, people around the world are no longer limited to the constraints of place and time to socialize with each other. The newly introduced concept of digital media has transformed the way our society function and socializing no longer requires the physical presence of society member. As a result of the introduction of multiple social media platforms, people from all over the world can now engage in local, national, and global events, participating in society and expressing themself in an open arena where physical boundaries do not stand in the way.

Today and after a very short time of experiencing the advantages of social media platforms, our society has become almost totally depending on such platforms, and most social events and happenings does not pass away from being recorded and discussed in the wide arena of social media. This effect generates a huge amount of valuable data that has a big potential of revealing the type and strength of social influence between society members, this opens from many useful applications that in multiple fields of society.

The most obvious application from social data is understanding how social media is used as a tool to mobilize groups of people in controversial social events such as political elections. The 2016 presidential election in USA and the allegation of Russian interference is one example that can benefit from analyzing social influence on social media.

Furthermore, by mapping and visualizing social influence between social media users, we can speed up and improve the detection of fake news and other illegal activities on social media. By removing such damaging effect to any social environment, we can create a better and more healthy society that benefits all its members.

Social influence is also highly valuable for commercial use, as many companies are interested in detecting different types of social influence to reveal new marketing trends and allow businesses to develop more specialized marketing strategies and customized products which often increase the competition in economy and generate more values in society.

These were some applications that benefit from analysis of influence in social media, and there is still both uncovered and undiscovered areas where understanding social influence is highly crucial for the purpose of the application.

# Introduction

Data from social media has a great potential in revealing the strengths of influences between network users, just like in real life every action a user commit and how society members react to this action can serve as an object for analysis which helps in drawing a big but rather detailed picture of how users influence each other across many societies and fields.

The aim of this research is establishing a ground foundation for extracting information about user activities on social media and use such information to detect social influence between network users. Such foundation is desired to make up the core of a future technical solution that enables social media analyzers with little or no technical experience in data processing and visualization to perform social media analysis on regular periods with a continuous timeline.

To serve this purpose, we start by determining the common characteristics in available functionalities between the most popular social media platforms, then produce a data model based on similarity between them. This data model is crucial to guarantee the reliability and flexibility of algorithms and technical systems that builds on top of this research; both reliability and flexibility are two sides of the same coin, as when basing the data model of the system on similarities between data provided by as many social media platforms as possible, we naturally widen the range of analysis potentials of any applying digital platform, making the solution more flexible and highly reliable and adaptable for social media analyzers.

After establishing an agreement on the data model to be used for collecting and storing crawled data from social media, we dive into the main core functionality of detecting social influence between network users. Multiple techniques of detecting such influences will be implemented to fit the different needs and purposes of the final application of analysis, regardless if the purposes are independent or can be tighten together in combinations. The desired result is a user influence graph where each node represents a participating user, while each edge between two given users representing the influence between these two users with respect to direction of influence and holding the strength and area of the influence whether it is in sport, politic, economy etc.

Following the previous effort, we evaluate the performance of the user graph model and go through test results from both dummy and real-life data using crawled data from a rising social media platform called “Reddit”. We will try to highlight the most interesting and useful features of the produced user graph and push its power of detecting influencers and their area of influence to the limit. The final test and evaluation process is a vital and necessary step to rely on the quality of the produced user graph model in any future analysis and technical solution.

This was a brief introduction of the upcoming research in a nutshell, but first let’s go through some interesting attempts in revealing user influence from social media.

# Related Works

Among the community of computer science, a wide variety of studies has focused on extracting information from the available social media platforms, and a big amount of effort has been dedicated to reveal influences between users to better understand the behaviour of individuals for many both commercial and non-commercial purposes. Research of social influence takes different forms and vary in size and scope, while some researchers take on the very fundamentals of detecting social influence others dive through to reveal specific influences and hidden behaviour patterns on different levels.

## Detecting and measuring user influence in social media

A good fundamental approach is described by a social network analysis carried out by Y. Guo, J. Cao & W. Lin. The fellow researchers are dividing the influence evaluation models into 2 main categories; the first category is based on network topology which measure social influence of a certain user by considering his degree, shortest path, and random walk characteristics, while the second category bases the influence between users on their interactions through different activities organized in tree-like structures like submissions and multilevel comments. However, and despite the reasonably good classification and overview these researchers offers, their paper lacks some proven results of an experimental approach [1].

## Data-driven Influence Learning

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

A short but rather interesting experimental and mathematical approach is introduced by a paper on Data-driven Influence Learning in Social Networks published by F. Wang, W. Jiang, G. Wang & D. Xie. In this paper, the process of influence diffusion is divided into two parts: the launcher (influence strength) and the receiver (influence threshold) which can generate an accurate and finer grained influence diffusion model according to this research. Furthermore, the researchers highlight the importance of having a solid criterium when scoring the strength and threshold properties of social influences. Another important acknowledgment is the difficulty and complexity associated with detecting influence relationships between users as a by-product of big datasets that usually include a considerable amount of noisy or less important datapoints, making it essential for any algorithm used in learning and testing the influence models to perform a minimal scan over the data in the most efficient way possible. [2]

## Gathering data and Crawling alternatives

Most well-known providers of social media platforms offer

developers and data scientist multiple endpoints and ways to extract data from their platforms for development or analysis. Some research spots the light on this initial aspect of gathering data from social media platforms. One significant research is one that mainly describe the alternative of Pushshift Reddit Dataset by J. Baumgartner, S. Zannettou, B. Keegan, M. Squire and J. Blackburn. [3] The research paper offers an undirected but also claimed to be a more efficient and flexible way to gather data from the “Reddit” social platform than by using the official Reddit API endpoint.

It also gives an excellent brief description of the FAIR data principles which is highly relevant when choosing the source of data especially when it comes to accessibility and findability.

Another advantage of this research is its extension in discussing a series of the other major alternatives for gathering data from Reddit, highlighting their strengths and weaknesses in a constructive manner.

## Classification of topics in social media platforms

As mentioned in the introduction, we are set to determine

the category of a detected influence between users, this opens up for the use of artificial intelligence for the purpose of classification between different topics where a certain user activity might fit in. In a research about annotating and detecting topics in social media forum and modelling the annotation to derive directions carried out by B. Athira, J. Jones, S. M. Idicula, A. Kulanthaivel and E. Zhang, a practical case study from an online health community was represented to give a good introduction of data pre-processing and cleaning, preceding to construct a reasonable mathematical approach for training and testing of a produced classification machine learning model.

Furthermore, the research offers a solution for a much-needed ability to minimize the amount of training data and dealing with the negative effects of label imbalance in a training dataset.

Another contribution of this research is the use of various deep learning algorithms to classify posted content such as CNN, LSTM and BiLSTM, all in which enable the researchers to achieve a promising F1-score of about 0.75 to 0.80 in topic classification.

The research has an excellent and well-performed evaluation and testing phase, where metrics of evaluation are carefully examined and explained in a good scientific approach [4].

## Choosing a study case social media platform

Determining which social media platform to crawl for testing and evaluation purposes is an important choice in the path to producing a user influence model that is flexible and useable in as many social media platforms as possible, this is why it is desirable to base the real life study case on actual real life datasets from a digital media platform that shares common user functionalities with as many popular social media platforms as possible, examples of such functionalities are post, submissions, comments, and upvotes or commonly known as likes.

A social media platform that satisfies all these user functionalities is “Reddit” which is examined by the research called “Information and Social Analysis” carried out by T. Steinbauer at the University of California, Santa Barbara. [5] Steinbauer starts off with a brief but very constructive comparison between the most popular social news sites with Reddit included. The core of Steinbauer’s research lays in his analysis of subreddits, submissions and comments on Reddit which can serve as a foundation prove for why Reddit should be used in evaluating the performance of a user influence graph model and its ability to view the most influencing users in a social media platform, the reason behind this is that Steinbauer gives a detailed analysis on which subreddits seems to have the most of activity and also precedes in building into constructing a user graph model that helps showing which user has the highest influence based on users interactions through comments.

However, submission authors are not included in the dataset of the constructed user graph, making the user graph less reliable when ignoring the often-significant role of posters in generating discussions on social media. Another downside of Steinbauer’s user graph model is not using any other criteria than user interaction through comments, such as the upvote score or number of thread comments to a certain comment or submission.

Although Steinbauer has introduced a detailed overview of his evaluations and analysis’s results, there is still a big question mark on the technical details as no algorithms for constructing the user graph has been presented in detail.

# Chossing a case study social media platform to crawl for test and evaluation

There exist a wide variety of popular social media platforms and most of them are constantly gaining popularity among users from all over the world, the following figure below shows the top 11 most popular social media platforms in the last decade in the days from the beginning of 2011 to the end of 2020. []



Although Facebook is the definite leading social media platform, but there exists a competition in popularity when looking at the next top 10 platforms, with Reddit located in the middle of the popularity overview. A normal side effect of a more popular social media is the large amount of data users generate on such platforms making the platform slow to crawl and extract data although data from a more popular media often has a higher integrity. To keep a balance between data integrity and easiness in findability and accessibility, we will try to compensate between these 2 factors by choosing a medium popular social media platform for testing and evaluation purposes. Reddit is located at the middle of the popularity range in the figure above and can line up as a candidate in a study case for development, testing and evaluation purposes in this research.

Many of the most popular social media platforms tends to specialize in a certain area or field of social activities such as LinkedIn for professional life, and Facebook on the other hand mostly used for private and personal socializing, some digital platforms combine aspects from both areas such as the so-called digital news platforms that offers its users an opportunity to interact with each other in many aspects of socializing like professional and personal life combined. Reddit is considered as one those digital news platforms which is still gaining popularity and increasing in content since its launch in 2005. []

A user on Reddit can create or join a group, make a submission on any group and comment on any submission or comment of other users. A user can join a group, but it is not obligatory to join a group to be active in them or read their content, these groups tend to specialize in a certain topic of interest in society and for many users it is seen in a way that is somehow like reading the newspapers which is often divided into pages for multiple areas of concern such as politics, economy, or sports. The high separation between topics of interest in Reddit makes it this platform ideal for testing how well an influence detecting algorithm can discover and classify different types of influence between users.

Reddit differs from other social media platforms in the sense that Reddit attracts users by their interest in topics and events in their social surroundings, while other social media often relays on the social affiliation of a future user. However, and on the other hand, many other social media platforms share a lot of common user functionalities with Reddit, such as groups, submissions, and comments.

This high similarity between Reddit and most popular social media platforms along with Reddit’s ability to separate users into multiple different social groups makes Reddit very suitable as an evaluation study case for this research as common functionality increases the algorithms flexibility for future use on other social media platforms, and its separation of social environments in groups serves the purpose of comparing the predicted type of social influence between users to the actual definition of the group where the interaction between users has occurred to give us an idea of how well our model is classifying topics of social influence.

Although Reddit is a user-oriented platform, its users often prefer to be anonymous, which is useful when presenting results with having to worry about neutrality issues, but Reddit’s users also use a username that can be used to identify a person account with revealing their identity, which is a useful feature for general research and data analysis as we are working on this project.

Another good reason for choosing Reddit as a study case is the highly developed endpoint crawling API which is very object-oriented and offers a wrapper library for the Python language that takes off programmers the bother of dealing with http requests and latency issues as all of this is taken care off in the background of the Python Reddit API Wrapper. The Wrapper is free to use but it requires a registration which once done offers no restrictions on how often Reddit is crawled, unlike crawling by adding “.json” to the URL which have many downsides such as limitation for under 100 submissions at a time, and the blockage of multiple requests from the same IP address as a prevention measure from Reddit to stop denial of service attacks. All these downsides are escaped by using the Python Reddit API Wrapper which increases the reliability and stability of data streams from reddit. In other word the PRAW python module satisfies the following FAIR data principles:

* Findability:

Once using PRAW, it is easy to find and retrieve data from Reddit no matter how detailed the data is.

* Accessibility:

As mentioned earlier a programmer does not have to deal with http request and latency issues as when using a traditional API endpoint, this makes the programming experience much easier allowing programmers to focus on the objective of their work.

* Interoperability:

A good documentation and maintaining history of the PRAW module along with its popularity between programmers who are crawling Reddit gives it an excellent record of ability to integrate with different products and systems that uses it.

* Reusability:

PRAW is very object-oriented in both query language and retrieval results. This is very helpful for the usability for integration in different projects and technical solutions both in present and future technologies.

Based on the above four FAIR data principles and the user habits analysis between the most popular social media platforms, Reddit makes a good case study in the testing and evaluation process for us seeking to detect user influence and their area of influence, we shall than design our ground data structure to adapt for the common functionality between Reddit and the most popular social media platforms as we will go through in the upcoming section on data structures.

# Definiing a ground data structure

Flexibility of design is an important requirement of this research, as we aim to apply the influence detecting algorithms on as many social media platforms as possible, and although this might be difficult to achieve as a result of the wide variety of available social media platforms and their different user-functionalities, we can still notice some common user functionalities between the most popular social media platforms such as LinkedIn, Facebook and Reddit, this common functionality is no accident, as these social media platforms most likely inspired from real life social interactions to begin with, which in turn is a natural advantage for our application.

After studying the available user functionalities in Reddit compared with these same functionalities on the most popular social media platforms, it is easy to see a big potential for developing a generalized data model that can be used to structure data crawled from any of the applying social media platforms. It is there important to consider the desired results of this research before establishing a ground data model.

Social influence can be defined to be the ability of one society member to change the thoughts or behavior of another society member, and although this definition is simple, the complexity is hidden in the way social influence plays out in real life society. Some people get influenced without any big significant reaction that can be recorded and studied, such influence is said to be passive influence, an example of such influence is reading a newspaper where the reader gets influenced without adding any comments to the content.

The main goal of this research is to use recorded data from social media to visualize the influence flow between a group of people in a social interaction. For this reason, we are going to look at active social influence where we would expect the person who get influenced to react by submitting an activity on the content of influence. This requires an activity-based model, where activities such as submissions and comments are considered as indicators for social influence.

The second requirement of this research is the importance of visualizing the flow and direction of influence between members in a social media interaction. For this purpose, we will be building an interaction-based model that is able to retain the origin and target of each detected and measured influence, which benefits the storage of influence direction and in the big picture can be used to visualize the entire flow of social influence between society members.

The model can initially be based on four different entities a user can create and interact to; these entities are:

1. Network

Which holds information about the crawled social media platform, having this entity, makes it possible to study multiple social media platforms at the same time which increases the flexibility of design.

1. Group

A group contains a bundle of submissions posted by users of the group. It also contains information about a certain group in form of identification and other attributes such as the group ID and name.

1. Submission

A submission is posted by one user and is assumed to be in text format with the possibility of further extension to multiple other formats like images and links in the future. It has multiple useful information stored in its attributes such as the current number of- comments, and -upvotes on this submission, along with information about the author of this submission and other identity attributes.

1. Comment

The comment entity is very similar to the submission entity containing a body text, identification and author and location information. However, it has an additional feature which its ability to be a parent and/or a child of another comment. This means that comments can be modelled as a tree data structure that can grow unlimited.

The figure below shows a diagram of an entity-relationship model that will make the base of our data structure further on in this research. In addition to the four entities explained earlier in this section, four relationships bind these entities together defining their relations to each other. A network can contain multiple groups, and a group can contain multiple submissions or posts, where users can either comment on those submissions or on other comments that is a child descendant of the comment tree of a certain submission.



Most of the popular social media platforms contains the four identified entities in this ER-model, although they might have a different name, form or purpose such as a company page on LinkedIn or a user profile on Facebook, both can be treated as groups just like Reddit groups as well.

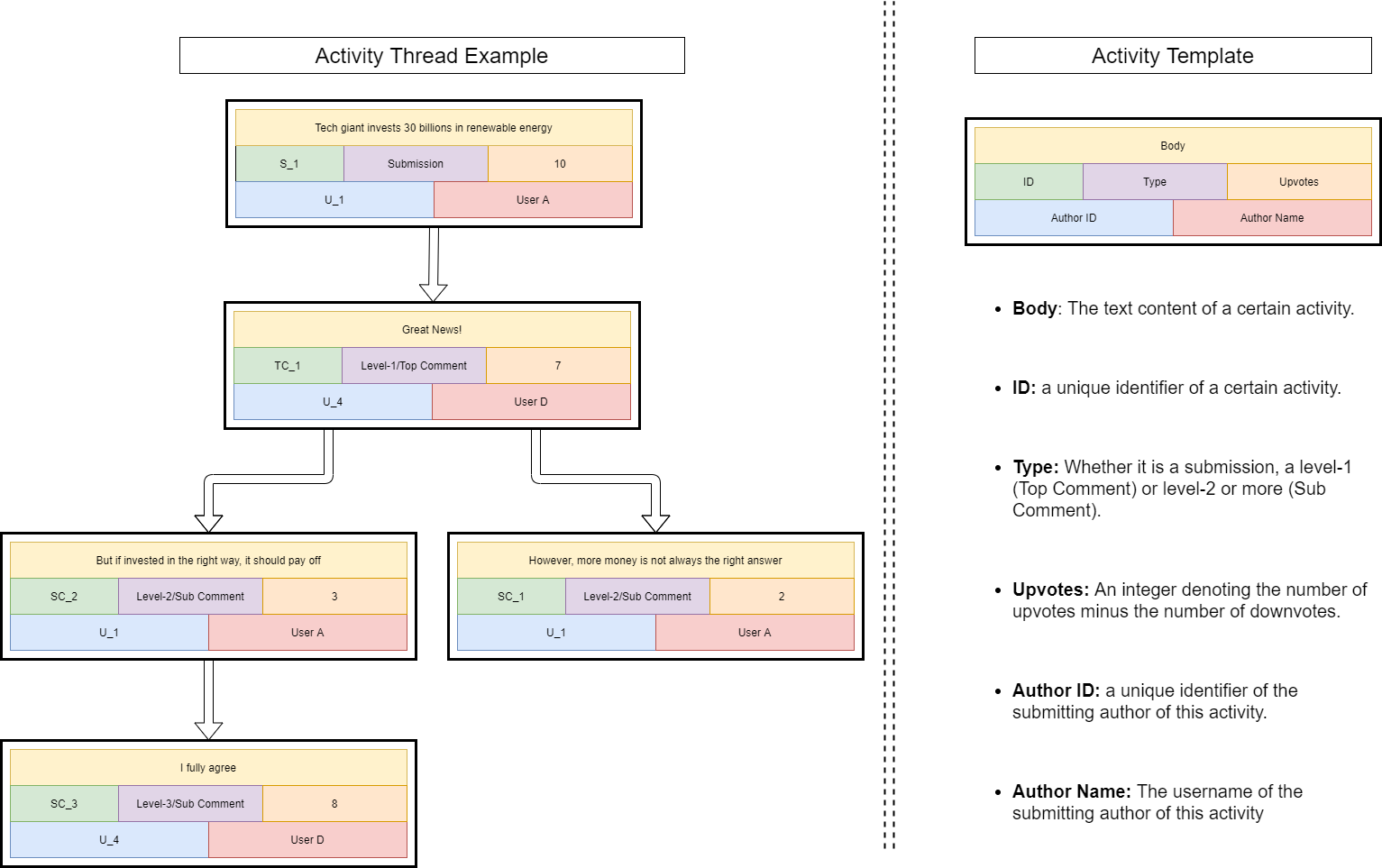
The attributes of entities open for more flexibility as we might have the need to extend or shrink our ER-model in the future i.e. by not including an upvote attribute or by adding a reaction attribute to submission and comments. But also, attributes are generalized to match the most available details about these common entities between the most popular social media platforms.

Now that we have a ground Entity-Relationship model to base our data structure on, we can proceed into discovering influences between users based of the interactions between them using these entities and relationship roles established in this section. In the next steps, we will elaborate the different stages of our method for detecting and extracting social influence on a social media platform, the algorithms used in this method have multiple dimensions to reveal the strength and type of influences between different members.

# USER ACTIVITY THREAD

Based on our previously established data model, we can observe a clear hierarchy between the following four subject entities: network, group, submission, and comment. This hierarchy enables us to model the data from a social network as a tree structure where network contains multiple groups that in turn contain a series of submissions which can contain multiple branches of comments, comments can have their own comments resulting in a tree that can grow endlessly as users add more comments on other previous comments.

The figure below shows a small example of just on submission and it is thread comments, we call this thread the user activity thread as each single node in this thread represent an activity that a certain user had performed, activities can be either a submission which we set to be the root of the activity thread tree, or a comment which can only have one parent that is to be a submission or an another comment, and it can have multiple children as users can comment on this comment.



In the next section, we will see how we can use this activity thread to give birth to a weighted and directed graph that shows the interaction flow between users performing these activities.

Since every tree is naturally a graph, we can benefit from our hierarchical approach to set up

As comments can have their own thread comments, we shall establish a categorization of comments based on their position in the hierarchy tree, where we can use 2 types of comments:

1. Top Comments: comments made directly on a submission.
2. Sub Comments: comments made on top comments or other sub comments.

The distribution of comments between top comments and sub comment can give a picture of how involved the network users are in each submission, as more sub comments than top comments can indicate that users are talking a time to read and react to top comment in addition to the submission which show a greater engagement of users than if users just keep themselves to writing top comments directly on the submission.

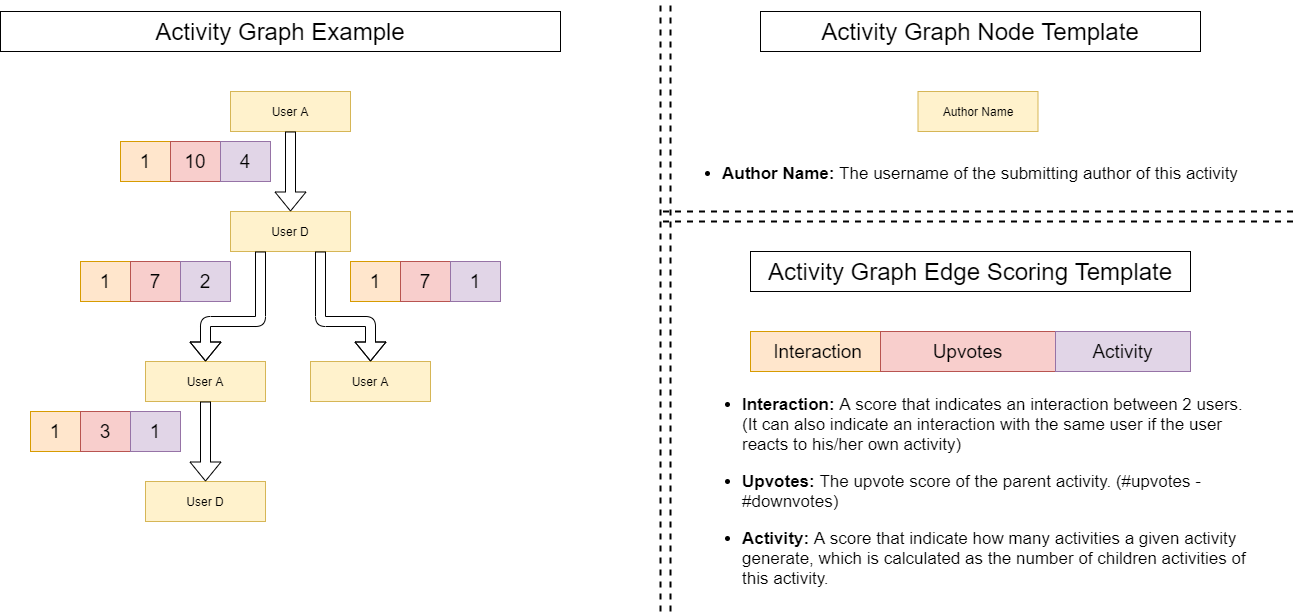
# USER ACTIVITY TRE & gRAPH

Since every tree is naturally a graph, we can take benefit from our hierarchical activity thread to construct a directed graph where each edge in this graph is directed from the parent activity to one of its child activities. This edge can represent an interaction between 2 users where target and source nodes are represented by their author, showing that the target user has reacted to an activity submitted by the source user, which in turn can be considered as an influence indication from a source user to a target user connected by a directed influence edge.

But knowing the direction of influence is not enough, as we wish to grade all known influences between users, so we can determine how strong a certain influence is in comparison to other influences in the graph topology. For this reason, we should give each influence edge a score that shows how strong this influence is, many techniques can be used here for scoring influences, however, each technique have its own strength and weaknesses and the analysis system can only get stronger by offering multiple scoring techniques to be used in analysis, in this project we have chosen 3 different scoring techniques represented in the table below, each with some of its strength and weaknesses.

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Explanation | Strengths | Weaknesses |
| Interaction | Measures how many times a user B has reacted to activities performed by user A. | High interaction score shows a clear strong influence between users. | In practice and depending on the network, it is very rare to detect multiple interaction between users as most users tends to interact one or two times in most cases. |
| Upvotes | Relay on the opinion of other reading users on the parent activity to score the influence from a user to another. | Gives a democratic approach that enables us to know weather the influence is supported or downvoted by the majority of interested audience. | In some networks such as Facebook, it is not possible for users to downvote an activity, which leave us without knowing whether the influence is supported or not by the society as audience. However, it is still possible the count of upvotes as a measure where 0 is the lowest score of an influence. |
| Activity | Gives an influence its score by measuring how many activities descendants the parent activity has been able to generate overall. | As central people in society tends to influence the most people by generating a great number of activities such as politicians, this score can help reveal who are those central people and which users do they have an influence on. | A downside of this scoring technique is that all child activities will get the same score which is the number of descendants activities of the parent activity, this downside can be compensated when users are interact with each other in different comments and comment threads. |

The figure below contains a graph built on the skeleton of the previous activity thread. It converts nodes from representing an activity to represent the user of that activity, while preserving the hieratical structure as it is. In addition, the 3 scoring techniques are explained and visualized on each edge.



After establishing our practical definition of influence and its strength based on different measuring techniques, we can digest the structure of the activity graph to produce a user influence graph that have a unique node for each user and a unique directed edge from an influencer to the user being influenced. More details about how to produce and read the user influence graph is presented in the next section.

# uSER iNFLUENCE gRAPH

##### References

[1] Y. Guo, J. Cao and W. Lin, "Social Network Influence Analysis," 2019 6th International Conference on Dependable Systems and Their Applications (DSA), 2020, pp. 517-518, doi: 10.1109/DSA.2019.00093. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] F. Wang, W. Jiang, G. Wang and D. Xie, "Data-Driven Influence Learning in Social Networks," 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), 2017, pp. 1179-1185, doi: 10.1109/ISPA/IUCC.2017.00177.

[3] Baumgartner, Jason & Zannettou, Savvas & Keegan, Brian & Squire, Megan & Blackburn, Jeremy. (2020). The Pushshift Reddit Dataset.

[4] Balakrishnan, Athira & Jones, Josette & Idicula, Sumam & Kulanthaivel, Anand & Zhang, Enming. (2021). Annotating and detecting topics in social media forum and modelling the annotation to derive directions-a case study. Journal of Big Data. 8. 10.1186/s40537-021-00429-7.

[5] T. Steinbaur, ‘‘Information and social analysis of Reddit,’’ inProc.TROYSTEINBAUER CS. UCSB. EDU, 2012, pp. 1–12. [Online]. Available:http://snap.stanford.edu/class/cs224w-2011/proj/tbower\_Finalwriteup\_v1.pdf

**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.**