COMP39/9900 24T3

Computer Science/IT Capstone Project

Project #17:

Finding, extrapolating, recommending and map connections in social media data sets

Members:

James Hanly
z5312477
z5312477@ad.unsw.edu.au

Justin Takayasu (*Scrum Master*)
z5363546
z5363546@ad.unsw.edu.au

Magan Leong
z5378443
z5378443@ad.unsw.edu.au

Michael Wang
z5358697
z5358697@ad.unsw.edu.au

Tsingying Xu
z5388553
tsingying.xu@student.unsw.edu.au

Ivan Fang (Product Owner)
z5418045
z5418045@ad.unsw.edu.au

Due Date: 9pm, Tuesday, 1st Oct 2024

1. Background

1.1 Problem Domain

In social media, users are naturally inclined to connect with others who share similar interests, mutual acquaintances, or common goals. They seek to expand their networks and discover like-minded individuals with whom they can form meaningful relationships. However, manually finding such connections is both challenging and time-consuming, especially given the sheer volume of active users. Moreover, many existing recommendation systems are generic, suggesting new users without clearly explaining why they might be a good match.

This project aims to solve this issue by developing an intelligent extraction and recommendation system tailored to each user. The extractor gathers key information about the user's interests, behaviours, and existing connections through their public post history. Using this data, the recommendation engine identifies potential connections, highlighting users who are most likely to share commonalities or mutual ties. The system then presents a curated list of recommended users along with a visual map that clearly illustrates the reasons behind each suggestion, offering users deeper insight into how these connections were made.

1.2 Existing Systems

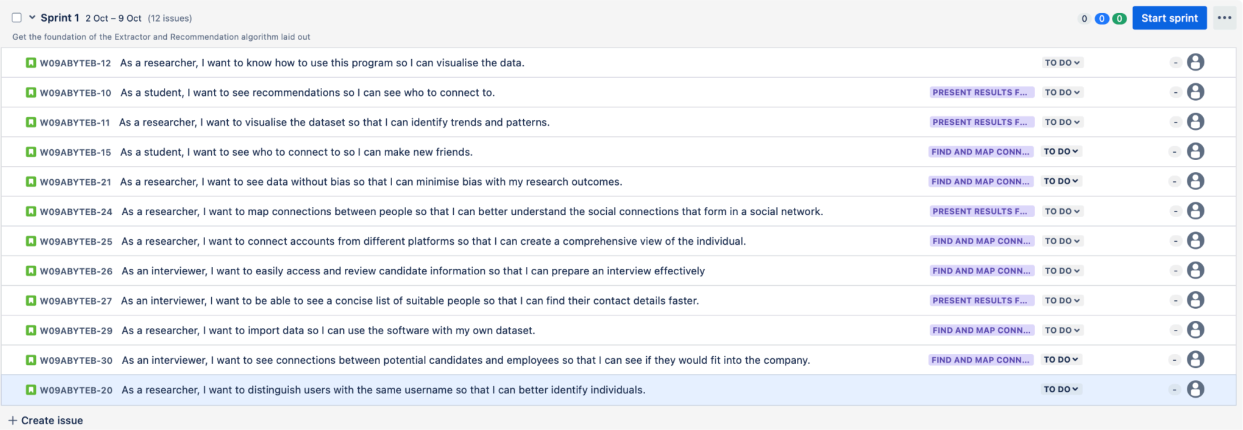
Facebook's "People You May Know" Suggestion Algorithm

The Facebook Friend Suggestion algorithm is one of the most used popular recommendation algorithms on a social media site today. It is a machine learning algorithm that uses content-based and collaborative filtering to detect connections between data sets and tries to predict what a user might like based on their prior interaction history. By taking predefined data such as age, gender and location, as well as interaction information like the amount of time spent on specific people's posts, it is able to generate a comprehensive comparison and suggest users that are most suitable. The usage of K-means clustering additionally allows for more precise grouping of users, enabling the algorithm to pick a recommendation from a smaller subset of similar users.

However, the specifics of the algorithm are made vague, and it is possible that Facebook is utilising additional data from individuals which should have been private. This inevitably led to the sparking of doubts and criticism from the general public about the safety of their privacy. In order to avoid this, we aim to use purely publicly available data or obtain consent from the individuals involved, as well as document our processes as transparently as we can.

2. User Stories and Sprints

2.1 Product Backlog



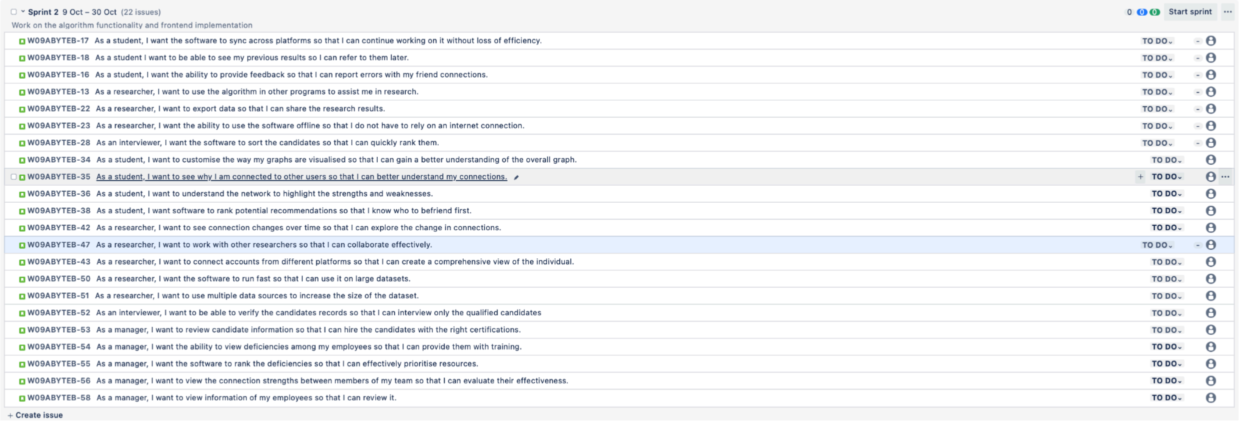
□ ⌄ **Sprint 1** 2 Oct – 9 Oct (12 issues) — 0 0 0 Start sprint ···

Get the foundation of the Extractor and Recommendation algorithm laid out

W09ABYTEB-12	As a researcher, I want to know how to use this program so I can visualise the data.		TO DO
W09ABYTEB-10	As a student, I want to see recommendations so I can see who to connect to.	PRESENT RESULTS F...	TO DO
W09ABYTEB-11	As a researcher, I want to visualise the dataset so that I can identify trends and patterns.	PRESENT RESULTS F...	TO DO
W09ABYTEB-15	As a student, I want to see who to connect to so I can make new friends.	FIND AND MAP CONN...	TO DO
W09ABYTEB-21	As a researcher, I want to see data without bias so that I can minimise bias with my research outcomes.	FIND AND MAP CONN...	TO DO
W09ABYTEB-24	As a researcher, I want to map connections between people so that I can better understand the social connections that form in a social network.	PRESENT RESULTS F...	TO DO
W09ABYTEB-25	As a researcher, I want to connect accounts from different platforms so that I can create a comprehensive view of the individual.	FIND AND MAP CONN...	TO DO
W09ABYTEB-26	As an interviewer, I want to easily access and review candidate information so that I can prepare an interview effectively	FIND AND MAP CONN...	TO DO
W09ABYTEB-27	As an interviewer, I want to be able to see a concise list of suitable people so that I can find their contact details faster.	PRESENT RESULTS F...	TO DO
W09ABYTEB-29	As a researcher, I want to import data so I can use the software with my own dataset.	FIND AND MAP CONN...	TO DO
W09ABYTEB-30	As an interviewer, I want to see connections between potential candidates and employees so that I can see if they would fit into the company.	FIND AND MAP CONN...	TO DO
W09ABYTEB-20	As a researcher, I want to distinguish users with the same username so that I can better identify individuals.		TO DO

+ Create issue

First Sprint



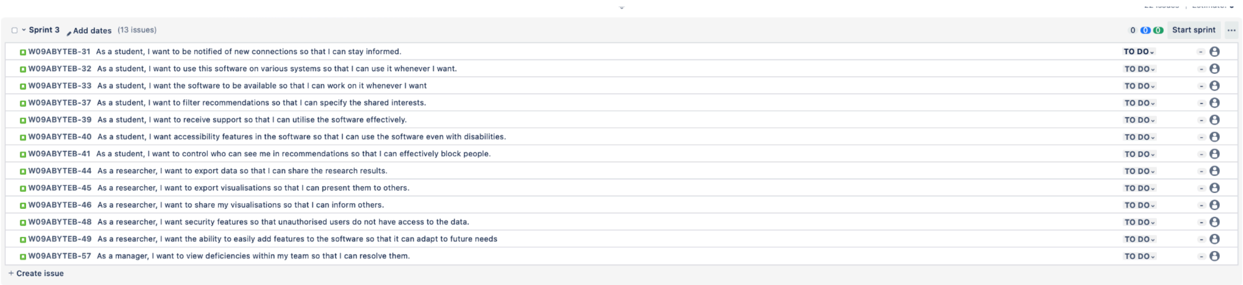
□ ⌄ **Sprint 2** 9 Oct – 30 Oct (22 issues) — 0 0 0 Start sprint ···

Work on the algorithm functionality and frontend implementation

W09ABYTEB-17	As a student, I want the software to sync across platforms so that I can continue working on it without loss of efficiency.	TO DO
W09ABYTEB-18	As a student I want to be able to see my previous results so I can refer to them later.	TO DO
W09ABYTEB-16	As a student, I want the ability to provide feedback so that I can report errors with my friend connections.	TO DO
W09ABYTEB-13	As a researcher, I want to use the algorithm in other programs to assist me in research.	TO DO
W09ABYTEB-22	As a researcher, I want to export data so that I can share the research results.	TO DO
W09ABYTEB-23	As a researcher, I want the ability to use the software offline so that I do not have to rely on an internet connection.	TO DO
W09ABYTEB-28	As an interviewer, I want the software to sort the candidates so that I can quickly rank them.	TO DO
W09ABYTEB-34	As a student, I want to customise the way my graphs are visualised so that I can gain a better understanding of the overall graph.	TO DO
W09ABYTEB-35	As a student, I want to see why I am connected to other users so that I can better understand my connections.	TO DO
W09ABYTEB-36	As a student, I want to understand the network to highlight the strengths and weaknesses.	TO DO
W09ABYTEB-38	As a student, I want software to rank potential recommendations so that I know who to befriend first.	TO DO
W09ABYTEB-42	As a researcher, I want to see connection changes over time so that I can explore the change in connections.	TO DO
W09ABYTEB-47	As a researcher, I want to work with other researchers so that I can collaborate effectively.	TO DO
W09ABYTEB-43	As a researcher, I want to connect accounts from different platforms so that I can create a comprehensive view of the individual.	TO DO
W09ABYTEB-50	As a researcher, I want the software to run fast so that I can use it on large datasets.	TO DO
W09ABYTEB-51	As a researcher, I want to use multiple data sources to increase the size of the dataset.	TO DO
W09ABYTEB-52	As an interviewer, I want to be able to verify the candidates records so that I can interview only the qualified candidates	TO DO
W09ABYTEB-53	As a manager, I want to review candidate information so that I can hire the candidates with the right certifications.	TO DO
W09ABYTEB-54	As a manager, I want the ability to view deficiencies among my employees so that I can provide them with training.	TO DO
W09ABYTEB-55	As a manager, I want the software to rank the deficiencies so that I can effectively prioritise resources.	TO DO
W09ABYTEB-56	As a manager, I want to view the connection strengths between members of my team so that I can evaluate their effectiveness.	TO DO
W09ABYTEB-58	As a manager, I want to view information of my employees so that I can review it.	TO DO

+ Create issue

Second Sprint



□ ⌄ **Sprint 3** Add dates (13 issues) — 0 0 0 Start sprint ···

W09ABYTEB-31	As a student, I want to be notified of new connections so that I can stay informed.	TO DO
W09ABYTEB-32	As a student, I want to use this software on various systems so that I can use it whenever I want.	TO DO
W09ABYTEB-33	As a student, I want the software to be available so that I can work on it whenever I want	TO DO
W09ABYTEB-37	As a student, I want to filter recommendations so that I can specify the shared interests.	TO DO
W09ABYTEB-39	As a student, I want to receive support so that I can utilise the software effectively.	TO DO
W09ABYTEB-40	As a student, I want accessibility features in the software so that I can use the software even with disabilities.	TO DO
W09ABYTEB-41	As a student, I want to control who can see me in recommendations so that I can effectively block people.	TO DO
W09ABYTEB-44	As a researcher, I want to export data so that I can share the research results.	TO DO
W09ABYTEB-45	As a researcher, I want to export visualisations so that I can present them to others.	TO DO
W09ABYTEB-46	As a researcher, I want to share my visualisations so that I can inform others.	TO DO
W09ABYTEB-48	As a researcher, I want security features so that unauthorised users do not have access to the data.	TO DO
W09ABYTEB-49	As a researcher, I want the ability to easily add features to the software so that it can adapt to future needs	TO DO
W09ABYTEB-57	As a manager, I want to view deficiencies within my team so that I can resolve them.	TO DO

+ Create issue

Third Sprint

2.2 User Stories (First Sprint)

As a researcher, I want to know how to use this program so I can visualise the data.

- The software must have a well documented “README” file and man pages explaining how to use the software, tweak settings, and supply / modify the user database

As a student, I want to see recommendations so I can see who to connect to.

- The software should present results in a readable manner.

As a researcher, I want to visualise the dataset so that I can identify trends and patterns.

- Friend recommendations must be visualised in an understandable graph that shows a clear overview of the algorithm output.

As a student, I want to see who to connect to so I can make new friends.

- The software should present recommendations tailored to each user.

As a researcher, I want to see data without bias so that I can minimise bias with my research outcomes.

- The software should present the data in an unbiased manner.

As a researcher, I want to map connections between people so that I can better understand the social connections that form in a social network.

- Algorithm output must be formatted in a standardised way, allowing for 3rd party software to request and handle the produced data.

As a researcher, I want to connect accounts from different platforms so that I can create a comprehensive view of the individual.

- The algorithm must be flexible enough to work with a range of data available from different social media platforms.

As an interviewer, I want to easily access and review candidate information so that I can prepare an interview effectively

- The software should make Information about the recommended candidates accessible.

As an interviewer, I want to be able to see a concise list of suitable people so that I can find their contact details faster.

- The software should provide recommendations in a neat list.

As a researcher, I want to import data so I can use the software with my own dataset.

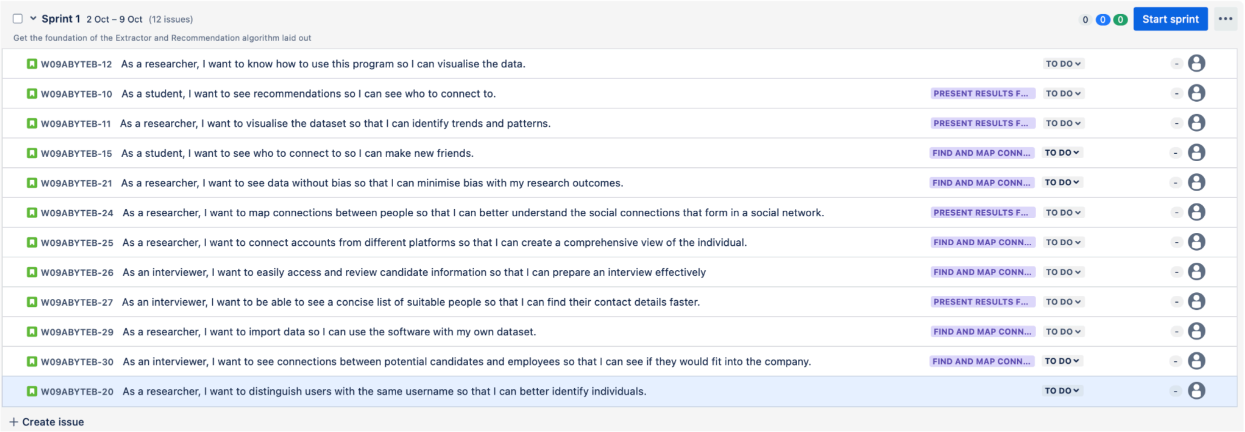
- The software should be capable of importing data in various formats.

As an interviewer, I want to see connections between potential candidates and employees so that I can see if they would fit into the company.

- The software should present relationships between members of different groups.

As a researcher, I want to distinguish users with the same username so that I can better identify individuals.

- The algorithm must be able to confidently link user profiles across social networks using available attributable and behavioural data.



□ Sprint 1 2 Oct – 9 Oct (12 issues)

Get the foundation of the Extractor and Recommendation algorithm laid out

W09ABYTEB-12	As a researcher, I want to know how to use this program so I can visualise the data.		TO DO	-
W09ABYTEB-10	As a student, I want to see recommendations so I can see who to connect to.	PRESENT RESULTS F...	TO DO	-
W09ABYTEB-11	As a researcher, I want to visualise the dataset so that I can identify trends and patterns.	PRESENT RESULTS F...	TO DO	-
W09ABYTEB-15	As a student, I want to see who to connect to so I can make new friends.	FIND AND MAP CONN...	TO DO	-
W09ABYTEB-21	As a researcher, I want to see data without bias so that I can minimise bias with my research outcomes.	FIND AND MAP CONN...	TO DO	-
W09ABYTEB-24	As a researcher, I want to map connections between people so that I can better understand the social connections that form in a social network.	PRESENT RESULTS F...	TO DO	-
W09ABYTEB-25	As a researcher, I want to connect accounts from different platforms so that I can create a comprehensive view of the individual.	FIND AND MAP CONN...	TO DO	-
W09ABYTEB-26	As an interviewer, I want to easily access and review candidate information so that I can prepare an interview effectively	FIND AND MAP CONN...	TO DO	-
W09ABYTEB-27	As an interviewer, I want to be able to see a concise list of suitable people so that I can find their contact details faster.	PRESENT RESULTS F...	TO DO	-
W09ABYTEB-29	As a researcher, I want to import data so I can use the software with my own dataset.	FIND AND MAP CONN...	TO DO	-
W09ABYTEB-30	As an interviewer, I want to see connections between potential candidates and employees so that I can see if they would fit into the company.	FIND AND MAP CONN...	TO DO	-
W09ABYTEB-20	As a researcher, I want to distinguish users with the same username so that I can better identify individuals.		TO DO	-

Screenshot of First Sprint

2.3 Schedule of Sprints

Sprint	Schedule
1	Wed 2 Oct 00:00 -> Wed 9 Oct 23:59 Goal: Lay out the foundational infrastructure of the project, short and sweet sprint
2	Wed 9 Oct 00:00 -> Wed 30 Oct 23:59

	Goal: Work on the algorithm functionality and frontend implementation, bulk of the work
3	Wed 30 Oct 00:00 -> Wed 13 Nov 23:59 Goal: FInalise and fine-tune code, focus on debugging, focus on wrapping up project

3. Technical Design

3.1 System Architecture Design

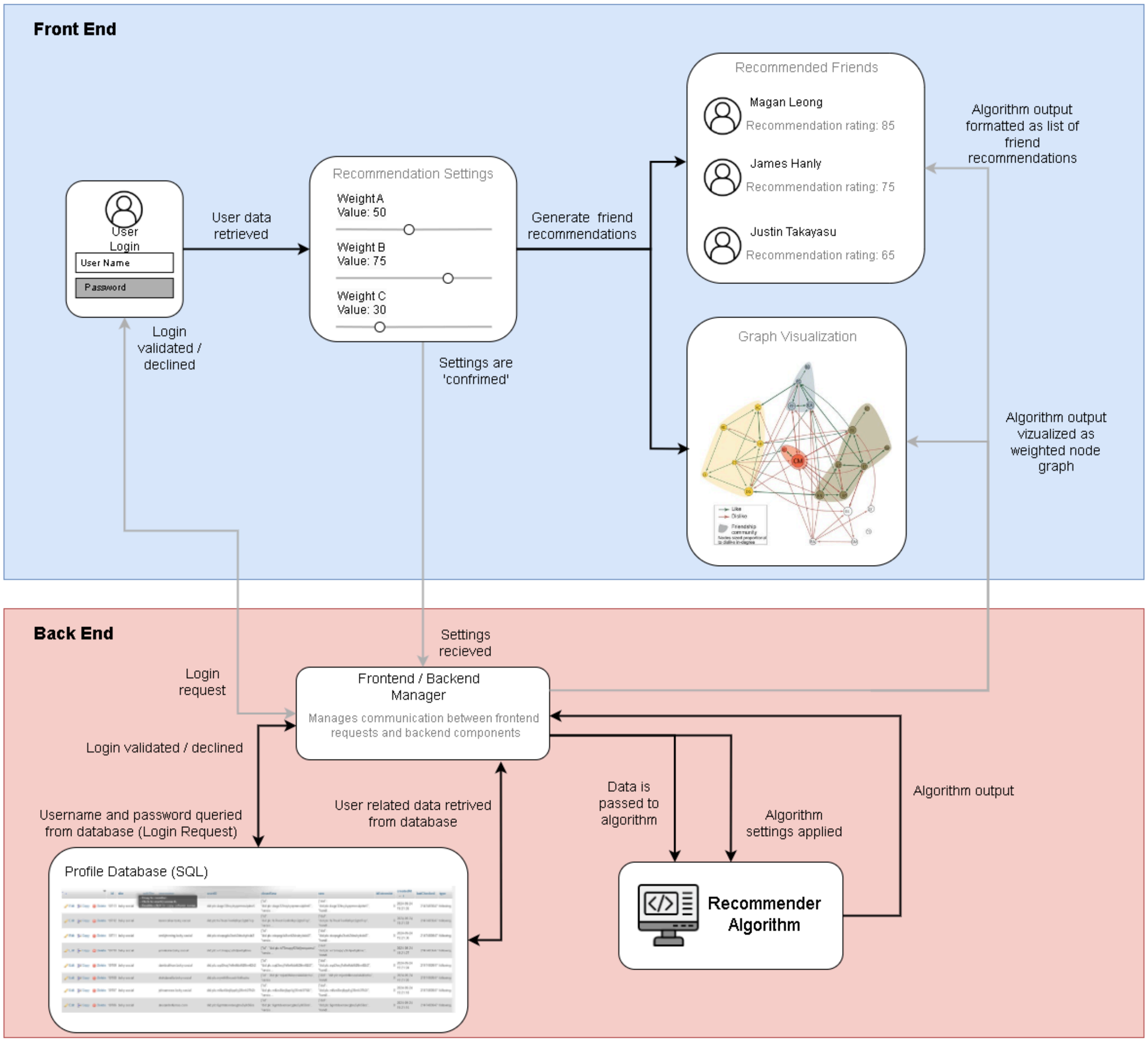

Front End
User Login
User Name
Password
User data retrieved
Recommendation Settings
Weight A
Value: 50
Weight B
Value: 75
Weight C
Value: 30
Generate friend recommendations
Recommended Friends
Magan Leong
Recommendation rating: 85
James Hanly
Recommendation rating: 75
Justin Takayasu
Recommendation rating: 65
Algorithm output formatted as list of friend recommendations
Graph Visualization
Algorithm output vizualized as weighted node graph
Login validated / declined
Settings are 'confrimed'
Back End
Login request
Settings recieved
Frontend / Backend Manager
Manages communication between frontend requests and backend components
Login validated / declined
Username and password queried from database (Login Request)
User related data retrived from database
Data is passed to algorithm
Algorithm settings applied
Algorithm output
Profile Database (SQL)
Recommender Algorithm

Fig 1: System Architecture Diagram, [visualised SN graph source]

3.2 Literature Reviews

In the digital age, establishing connections between any two people has never been easier: The surplus of social media platforms means users are more involved with trends and each other, connecting more users from drastically different interests and backgrounds. Establishing user connections within these networks is a central aspect of enhancing user experience, fostering community, and driving engagement. To address the challenges of effectively recommending new social connections, researchers have explored various methodologies, drawing upon data mining, algorithmic approaches, and the integration of multi-network information. This literature review examines recent contributions to the field, highlighting emerging trends, revolutionary approaches, and limitations associated with recommending user connections on social media.

["*Social Friend Recommendation Based on Multiple Network Correlation*" - *Shangrong Huang; Jian Zhang; Lei Wang; Xian-Sheng Hua*]

In this paper, the authors propose that the social aspects of a user plays an important part in the decision-making to becoming friends with other users, summarised together as a category called a "social role". These roles form different networks that are correlated to each other, known as contact networks and tag networks. Contact networks are based on whether users are friends, while tag networks are connected by important keywords that define the user. Through network alignment, features are selected based on the alignment of the two networks and important feature sets are constructed for each user. If two users are found to have a strong similarity in the tag network, there is a higher probability of these two users having a possible relationship and should be recommended to each other. The authors believe that by finding the correlations between these networks, they would be able to provide a more comprehensive and tailored friend recommendation system.

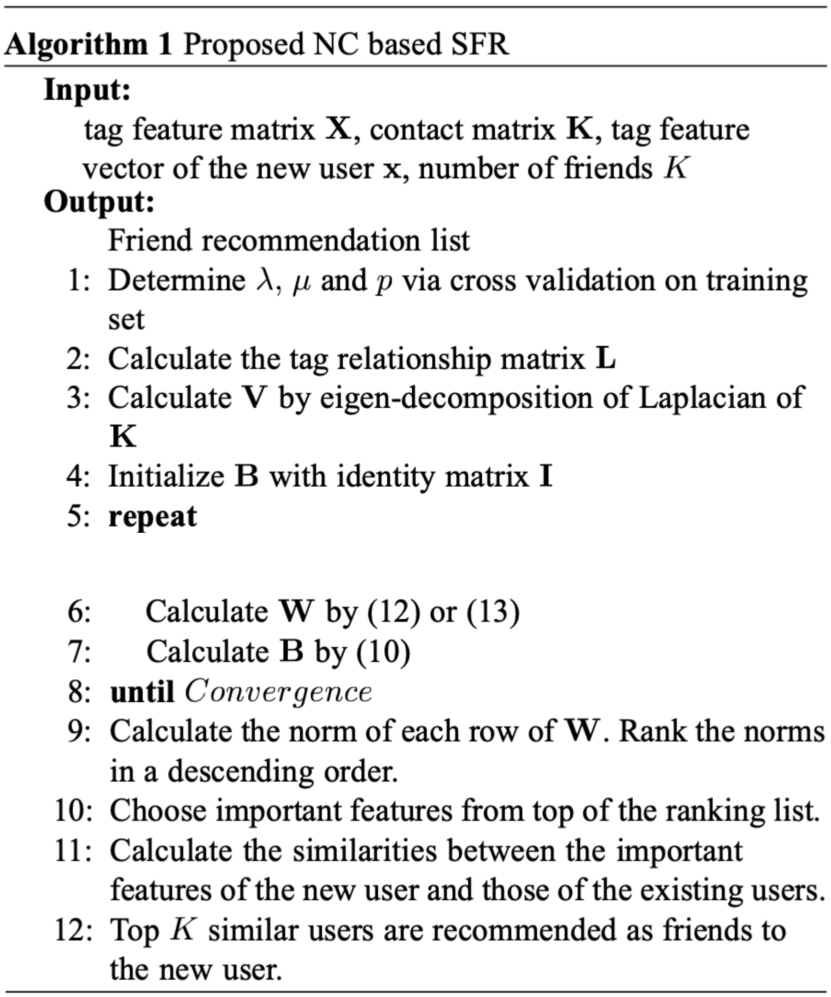
The algorithm uses Flickr as the main social media example, converting the site's tag words into "features". Each user's feature set may contain tags, photos, comments, geo-information, etc. However, only a small subset of that set contain features that correlate to an individual's friend-making decision. Considering two networks, we select features that correlate the networks well and preserve the original structure of the networks. These features are then constructed into a feature matrix using the TF-IDF method. Using this feature matrix, the user-user similarity between the networks can be minimised.

To ensure that structure is maintained after the gap minimisation, tag feature similarity measurements are gathered to represent the structure of the networks. By obtaining the relation scores of different features, a relation score matrix represented by L can be used to minimise the change in in-dimension distance via the following equation.

$$\min_{\mathbf{W}} \sum_{j=1}^{n}\sum_{i=1}^{n} \mathbf{A}_{1i}\mathbf{A}_{1j}\mathbf{L}_{ij} + \cdots + \sum_{j=1}^{n}\sum_{i=1}^{n} \mathbf{A}_{ri}\mathbf{A}_{rj}\mathbf{L}_{ij}$$
$$= \min_{\mathbf{W}} \operatorname{tr}(\mathbf{A}\mathbf{L}\mathbf{A}^T) = \min_{\mathbf{W}} \operatorname{tr}(\mathbf{W}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{W}). \quad (8)$$

Fig. 2. Formulation to minimise network deviation

Once the new graph network has been obtained, we compare the feature sets of the existing users to those of new users. The more similar the tag sets, the closer the two users should be. By ranking the tag similarity of these feature sets, we then choose the most similar members as potential connections to recommend to the new user.



Algorithm 1 Proposed NC based SFR

```
Input:
    tag feature matrix $\mathbf{X}$, contact matrix $\mathbf{K}$, tag feature
    vector of the new user $\mathbf{x}$, number of friends $K$
Output:
    Friend recommendation list
Determine $\lambda$, $\mu$ and $p$ via cross validation on training set
Calculate the tag relationship matrix $\mathbf{L}$
Calculate $\mathbf{V}$ by eigen-decomposition of Laplacian of $\mathbf{K}$
Initialize $\mathbf{B}$ with identity matrix $\mathbf{I}$
repeat
    Calculate $\mathbf{W}$ by (12) or (13)
    Calculate $\mathbf{B}$ by (10)
until $Convergence$
Calculate the norm of each row of $\mathbf{W}$. Rank the norms in a descending order.
Choose important features from top of the ranking list.
Calculate the similarities between the important features of the new user and those of the existing users.
Top $K$ similar users are recommended as friends to the new user.
```

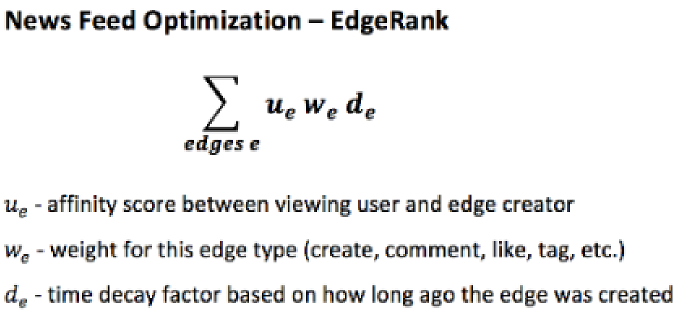
Fig. 3. Outline of Proposed Algorithm

The algorithm listed in this paper is notable for having a more in-depth recommendation system than simple content similarity matching, which is especially useful for providing precise results in large data sets such as the ones specified in our project. However, the experiments conducted by the paper have only considered the capabilities of two networks in detail. We will be adopting the theoretical solution provided in the paper for expansion to multiple networks. Furthermore, the paper uses Flickr as its only social network, which we obviously have to modify in order to fit a wider spectrum of social media.

[Understanding Social Media Recommendation Algorithms, 2021 - Arvind Narayanan]

In recent research academics have found that connections between users are established through recommendation algorithms which use information propagation, whether that be through subscriptions, user networks or algorithms, to bring relevant and engaging content to users. Professor Arvind Narayanan from Princeton University's paper, *Understanding Social Media Recommendation Algorithms*, delves into the complexity and vastness of social media platforms and the human-algorithm interactions that connect people over the world wide web. Connecting users to one other and recommending new trends is only possible through information propagation, and the algorithms that help propagate the right information to users that are most likely to engage with such content. As more users engage with similar content, connections between users begin to form and thus user networks are created: Narayan refers to data used by recommendation algorithms in social media platforms as signals, "Two people who have something in common—a hometown, a hobby, a community they are embedded in, a celebrity they follow—will both engage with posts that relate to that shared interests… Platforms vary in how much they place emphasis on this signal." (Narayanan, 2023).

Recommendation algorithms were first seen in companies such as Amazon, where users' purchase history influenced the items the website would recommend, and Netflix, where films and shows would be recommended to users based on their history and what users with similar histories were watching. Such recommendation algorithms can be seen to focus primarily on the business aspect of their platforms, aiming to increase revenue rather than encourage user communities. Observing recommendation algorithms in social media highlights the ability for such algorithms to build user connections and cultivate networks of people who share similar interests and views. Facebook's 2010 "EdgeRank" algorithm was a precursor to today's social media recommendation algorithms, optimising the news headlines a user would get recommended on their facebook feed by following a basic equation - observable from **figure 4** (Valentine and Wukovitz, 2013).



News Feed Optimization – EdgeRank

$$\sum_{edges\ e} u_e\, w_e\, d_e$$

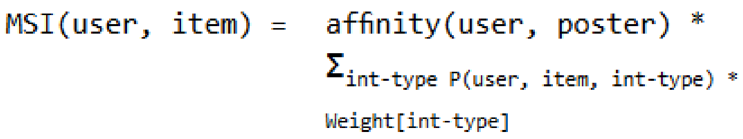
u_e - affinity score between viewing user and edge creator

w_e - weight for this edge type (create, comment, like, tag, etc.)

d_e - time decay factor based on how long ago the edge was created

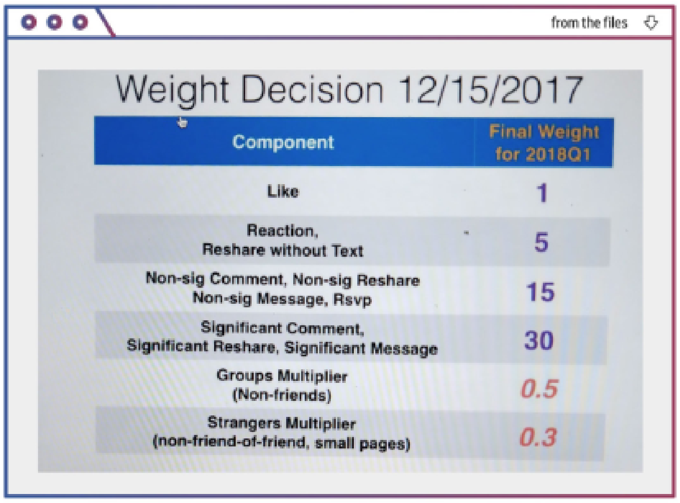
Fig. 4. The EdgeRank formula as presented at Tech Crunch in 2010 (Valentine and Wukovitz, 2013)

Facebook later replaced EdgeRank with their "Meaningful Social Interactions (MSI)" because EdgeRank based its output off manually inputted values by Facebook engineers, whereas the MSI algorithm learned and calculated scores through Facebook's data. This metric, of which a representation of the algorithm can be seen in **figure 5**, recommends content to users from highest MSI score to lowest. A potential method of calculating MSI scores can be seen in **figures 6 and 7**.



```
MSI(user, item) =  affinity(user, poster) *
                   Σint-type P(user, item, int-type) *
                   Weight[int-type]
```

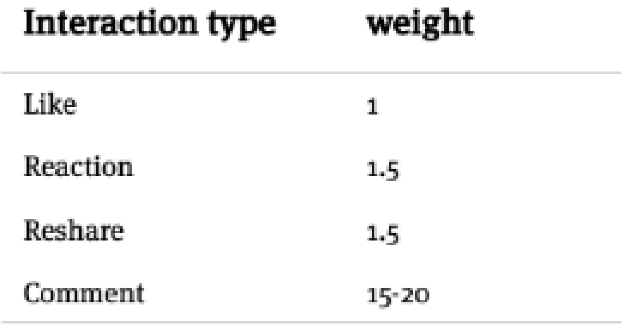
Fig. 5. MSI formula (Narayanan, 2023)



Weight Decision 12/15/2017

Component	Final Weight for 2018Q1
Like	1
Reaction, Reshare without Text	5
Non-sig Comment, Non-sig Reshare Non-sig Message, Rsvp	15
Significant Comment, Significant Reshare, Significant Message	30
Groups Multiplier (Non-friends)	*0.5*
Strangers Multiplier (non-friend-of-friend, small pages)	*0.3*

Fig. 6. Internal Facebook memo from December 2017 depicting a potential method of MSI score calculation (Narayanan, 2023)



Interaction type	weight
Like	1
Reaction	1.5
Reshare	1.5
Comment	15-20

Fig. 7. Interaction-type weights for the MSI formula in 2020 (Narayanan, 2023)

Though these recommendation algorithms yield mostly accurate and satisfactory results, they lack transparency towards their user base and so limit users' control over what they can and cannot see. Another limitation of such algorithms is the unpredictability of the virality of content on the platform.

Platforms emphasise feedback that is more frequent, i.e Youtube and TikTok observe both explicit - like and dislike buttons, comment sections - and implicit feedback - swiping faster or slower on videos, or swiping to view similar content - meaning that even if users aren't deliberately engaging less with specific content of a creator, the algorithm will reflect that. We can begin to address this problem by implementing a balanced scoring mechanism paired with a time decay function to accurately differentiate between the importance of old and new interactions. User feedback and transparency will also be crucial, helping to refine the scoring mechanism and recommender algorithm based on user preferences and trust.

["SocialMatching++: A Novel Approach for Interlinking User Profiles on Social Networks" - Hussein Hazimeh, Elena Mugellini, Omar Abou Khaled, Philippe Cudré-Mauroux]

As the popularity of social networks has grown rapidly over the last two decades, there is a growing need to accurately match a user's profile between social networks. This has become a complex problem, largely due to the rise of screen name duplication in social networks with many users. Hazimeh, Mugellini, Khaled and Cudré-Mauroux (2017) provide a novel approach, named "SocialMatching++" for interlinking a user's profile from one social network to another with higher precision than existing state of the art methods. SocialMatching++ (Match++) argues that reliable links between user profiles across platforms can be achieved in quantifying a user's personality. This is achieved through two novel methods, "DEscription Based Linking" (DEBL) and "Life Event Based Linking" (LEBL) which are based on the established approaches, "Attribute-Based" (AB) and "Content and Behavioural Based" (CBB) respectively.

The attribute based approach uses "basic profile attributes such as gender, age, location or profile images." When the attribute based method is applicable, a reliable profile link can be found by using a combination of profile attributes, weighted by importance. However, this approach is entirely dependent and therefore, limited by the quality and availability of the attribute data. This is seen in the method linking by use of detecting historical changes to profile attributes, where historical data is not often available and difficult to extract. The authors reference @I seek 'fb.me' as the baseline AB comparison as "it uses a variety of profile attributes" (p.11).

The content and behavioural based approach leverages user interactivity with a social network. This approach measures features such as posting rates, timestamps, geolocation and content of posts to quantify a user's personality, which can then be mapped across SN platforms. Among the CBB approaches outlined, HYDRA, is declared as "the most relevant and important contributions" (p.11) and is used as the CBB baseline comparison. This approach is limited by both the availability of content data and the sophistication in which behavioural patterns are extracted from said data.

The Match++ method was shown to be more effective than each CBB and AB approach. This was evaluated using precision as the evaluation metric against the corresponding baselines HYDRA and @I seek 'fb.me respectively.

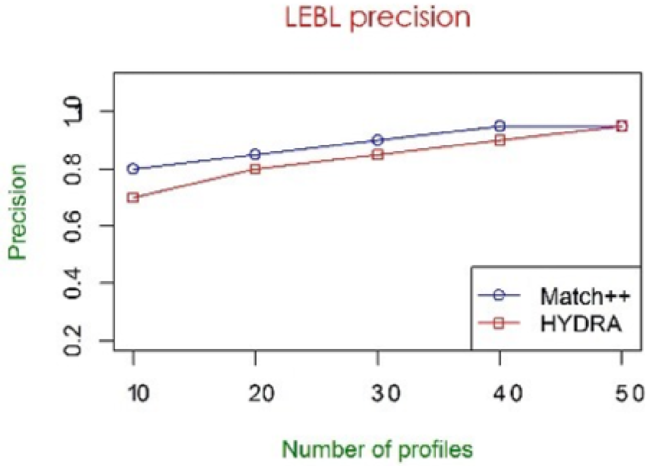


Fig. 8. LEBL precision compared to content-behavioural-based baseline

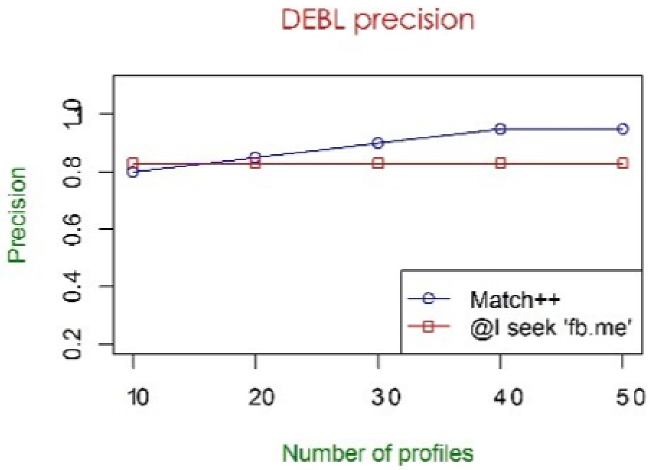


Fig. 9. DEBL precision compared to attribute-based baseline

The use of combining CBB and AB approaches in Match++ proves to be comparable to either baseline when compared individually, while allowing greater flexibility in the types of applicable user profiles. Furthermore, it is argued that because life event information is unique to a person, Match++ provides more accurate profile links than existing baselines.

However, this method isn't without limitation. The data required from user profiles to effectively apply LEBL isn't easily available on many social networks. The authors outline a need for Life events classification to become more sophisticated before linking methods that employ CBB can be used across a range of social networks.

["Recent developments in exponential random graph (p) models for social networks (2007)" - Ding Ma]*

Our project aims to identify, visualise, and recommend social media connections across various datasets and platforms. Relevant literature is grounded in social network analysis, particularly the Exponential Random Graph Models for understanding the structure of social networks. As highlighted in the attached paper, these models provide a sophisticated framework for analysing complex networks by capturing the relationships and structures within them. The traditional Markov random graphs were limited in their capacity to model social networks due to issues like near-degeneracy when applied to highly transitive networks (networks where friends of friends are likely to be friends). However, new specifications developed by Snijders et al. (2006) introduce higher-order statistics, such as alternating k-triangles and k-stars, which better capture transitivity and improve model fit.

Our project is focused on social media data, where novel techniques must extend these statistical models into the realm of cross-platform identity matching, data extrapolation, and recommendation systems. Our project can build on the literature in two ways. Firstly, Multiplatform User Matching. One novel approach is to apply graph models not just to one dataset but across multiple platforms, aiming to infer connections between users with similar attributes (e.g., usernames) but possibly varying platform behaviours. We could extend the ERGM methods by creating multilayer networks that integrate data across platforms. Secondly, Recommender Systems. For this aspect, the literature on collaborative filtering and content-based filtering methods can be enhanced by leveraging graph-based algorithms like Node2Vec to find recommendations based on structural similarity within the network.

There are three possible evaluation metrics. Firstly, goodness of fit. As discussed in the literature, ERGMs have used Monte Carlo maximum likelihood estimation to improve model fit compared to simpler likelihood estimations such as pseudo-likelihood. For our project, measuring how well the inferred networks match real-world data across platforms will be key. Secondly, accuracy and precision in user matching. Precision and recall, or the ability to correctly identify users across different platforms, will be essential metrics. Lastly, efficiency. Given the large datasets involved, evaluation should include time and space complexity, especially as the network size grows.

The first limitation is scalability. In our project, we can mitigate this by leveraging distributed computing frameworks like Apache Spark for large-scale social media data processing, allowing you to handle larger datasets more efficiently. The second limitation is data sparsity and near-degeneracy. Instead of purely relying on ERGM models, hybrid methods combining graph neural networks with traditional statistical models can be used. GNNs offer a data-driven approach to learning patterns in large-scale graphs while maintaining computational efficiency.

[Research challenges and opportunities in mapping social media and Big Data Ming-Hsiang Tsou]

The research talks mainly about two things. Firstly, Multi-scale spatiotemporal analysis. Combining geographic information system tools and social network analysis to map user behaviours. This methodology could be adapted to process user location data alongside user interaction data, thus adding a layer of location dynamics to recommendations. Secondly, Data fusion and linked data. Tsou's paper highlights integrating various Big Data layers to understand patterns. This is particularly valuable for the task of connecting users from different platforms, as it supports building robust connections across social networks. This could be built on by implementing machine learning techniques (like collaborative filtering or graph-based learning) that use location factors as features to improve recommendation accuracy, a novel extension not fully explored in existing literature.

There are three possible evaluation metrics. Firstly, goodness of fit. As discussed in the literature, ERGMs have used Monte Carlo maximum likelihood estimation to improve model fit compared to simpler likelihood estimations such as pseudo-likelihood. For our project, measuring how well the inferred networks match real-world data across platforms will be key. Secondly, accuracy and precision in user matching. Precision and recall, or the ability to correctly identify users across different platforms, will be essential metrics. Lastly, efficiency. Given the large datasets involved, evaluation should include time and space complexity, especially as the network size grows.

The first limitation is data sparsity. A common issue in Big Data and social media analysis is dealing with incomplete or sparse data, particularly for social connections. To overcome this, the system could implement graph completion techniques or predictive algorithms to infer missing links. Another limitation is noisy data. Social media data often contains irrelevant or noisy information. Developing noise-filtering algorithms or applying natural language processing for post content analysis can enhance the quality of the data used for recommendations. The last limitation is scalability. In our project, we can mitigate this by leveraging distributed computing frameworks like Apache Spark for large-scale social media data processing, allowing you to handle larger datasets more efficiently.

We have seen how establishing user connections on social media involves techniques like data mining, recommendation algorithms, and multi-network integration. Approaches such as multiple network correlation, social role-based algorithms, and advanced linking methods all emphasise the importance of combining diverse data sources—from user profiles to behavioural patterns—to enhance accuracy and personalization in social recommendations. However, challenges related to data scalability, transparency, and the unpredictability of user behaviour remain prevalent, requiring nuanced solutions.

3.3 Design Justifications

- Presenting recommendations
    - Presenting text
        - Raw text output (txt)
        - Tabulated data (e.g. csv, tsv, etc.)
        - Formatted reports (e.g. pdf, doc, etc.)
    - Graph visualisations
        - Image output (e.g. png, jpeg, etc.)
- Cross platform compatibility
    - Python backend
        - Static server side rendered web pages
        - Compliance with html standards
- State control, history tracking, progress sharing
    - Backend keeps track of state with separate database
- Scalability, High availability
    - Highly available services with redundancy
    - Scalable services
- Extensibility, Adaptability
    - Static analysis (e.g. static typing)
    - Runtime checking (e.g. assertions)
    - Exception / error handling
    - Strict adherence to style guidelines.
    - Use existing libraries if possible
- Import / export
    - Support common file types (e.g. csv, json, etc.)
- Feedback / support
    - Support ticket system
    - Feedback system
- Notification systems
    - Web based notifications, mobile notifications,
- Bias
    - The software should not distort data
- Security
    - User input sanitisation
    - Database validation

3.4 Project Objective Checklist

Project Objective/Functionality	User Story
Identifying suitable people through the recommendation system	As a student, I want to see recommendations so I can make new friends. As an interviewer, I want to be able to see a concise list of suitable people so that I can find their contact details faster.
Create connections between datasets of candidates that may be related	As a researcher, I want to map connections between people so that I can understand the structure of the network. As a researcher, I want to find and map people who have mutual connections so I can understand social circles. As a social media analyst, I want to discover direct and indirect connections between users so that I can identify social clusters.

Visualising the connections between candidates that led to the suggestions made	As a student, I want to visualise my own connections so that I can better understand my network. As an interviewer, I want to be able to verify the candidates records so that I can interview only the qualified candidates. As a researcher, I want to visualise the dataset so that I can identify trends and patterns. As a brand strategist, I want to be able to identify weak connections such that I can better estimate the user response to my brand.
Identifying and linking different accounts made by same user together	As a researcher, I want to connect accounts from different platforms so that I can create a comprehensive view of the individual.
Display information about recommended candidates	As an interviewer, I want to easily access and review candidate information so that I can prepare an interview effectively

References:

- Hill, K. (2017). *How Facebook Figures Out Everyone You've Ever Met*. [online] Gizmodo. Available at: https://gizmodo.com/how-facebook-figures-out-everyone-youve-ever-met-1819822691.
- Goswami, A. (2022). *Facebook Friend Suggestion Algorithm.* [online] Linkedin.com. Available at: https://www.linkedin.com/pulse/facebook-friend-suggestion-algorithm-arunav-goswami.
- Narayanan, A. (2023). Understanding Social Media Recommendation Algorithms. Academic Commons. [online] doi:https://doi.org/10.7916/khdk-m460.
- Valentine, A. and Wukovitz, L. (2013). Using The Filter Bubble to Create a Teachable Moment: A Case Study Utilizing Online Personalization to Engage Students in Information Literacy Instruction. Pennsylvania Libraries: Research & Practice, 1(1), pp.24–34. doi:https://doi.org/10.5195/palrap.2013.18.
- Huang, S., Zhang, J., Wang, L. and Hua, X.-S. (2016). Social Friend Recommendation Based on Multiple Network Correlation. IEEE Transactions on Multimedia, 18(2), pp.287–299. doi:https://doi.org/10.1109/tmm.2015.2510333.
- Zhang, Y., Tang, J., Yang, Z., Pei, J. and Yu, P.S. (2015). COSNET. Knowledge Discovery and Data Mining. doi:https://doi.org/10.1145/2783258.2783268.
- Hussein Hazimeh, Mugellini, E., Omar Abou Khaled and Philippe Cudré-Mauroux (2017). SocialMatching++: A Novel Approach for Interlinking User Profiles on Social Networks. PROFILES 2017.
- Hossain, I., Puppala, S., Alam, M.J. and Talukder, S., 2024. SocialRec: User Activity Based Post Weighted Dynamic Personalized Post Recommendation System in Social Media. *arXiv preprint arXiv:2407.09747*.
- Xu, J., 2022. Analysis of Social Media Algorithm Recommendation System. *Studies in Social Science & Humanities*, *1*(3), pp.57-63.
- Barragáns-Martínez, A.B., Costa-Montenegro, E., Burguillo, J.C., Rey-López, M., Mikic-Fonte, F.A. and Peleteiro, A., 2010. A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition. *Information Sciences*, *180*(22), pp.4290-4311.