Link Prediction and Cascading within Subreddits

Sunjeet Jena (ASU ID: 1218420294)

Ankit Sharma (ASU ID: 1219472813)

Oluwaseun Talabi(ASU ID: 1213104917)

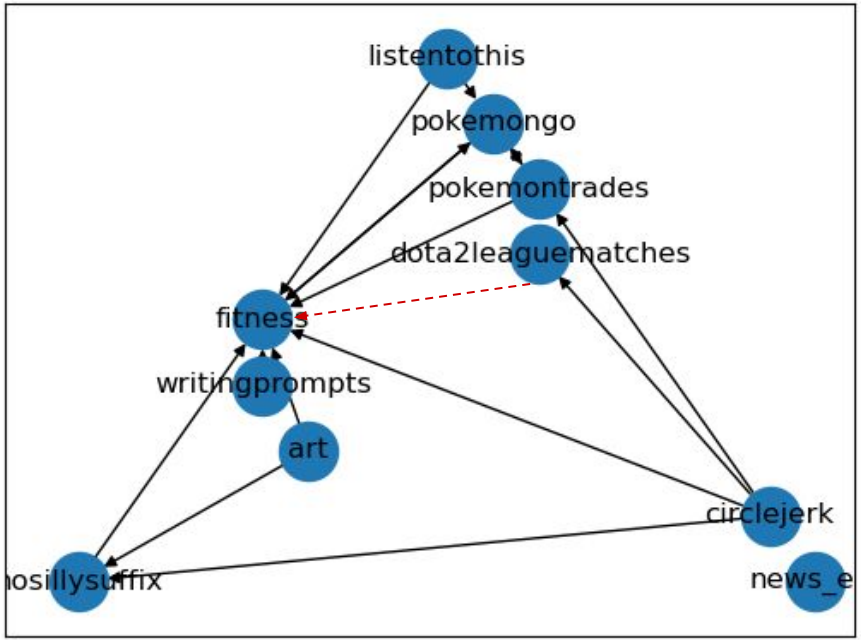
Akhil Mittal (ASU ID: 1219691005)

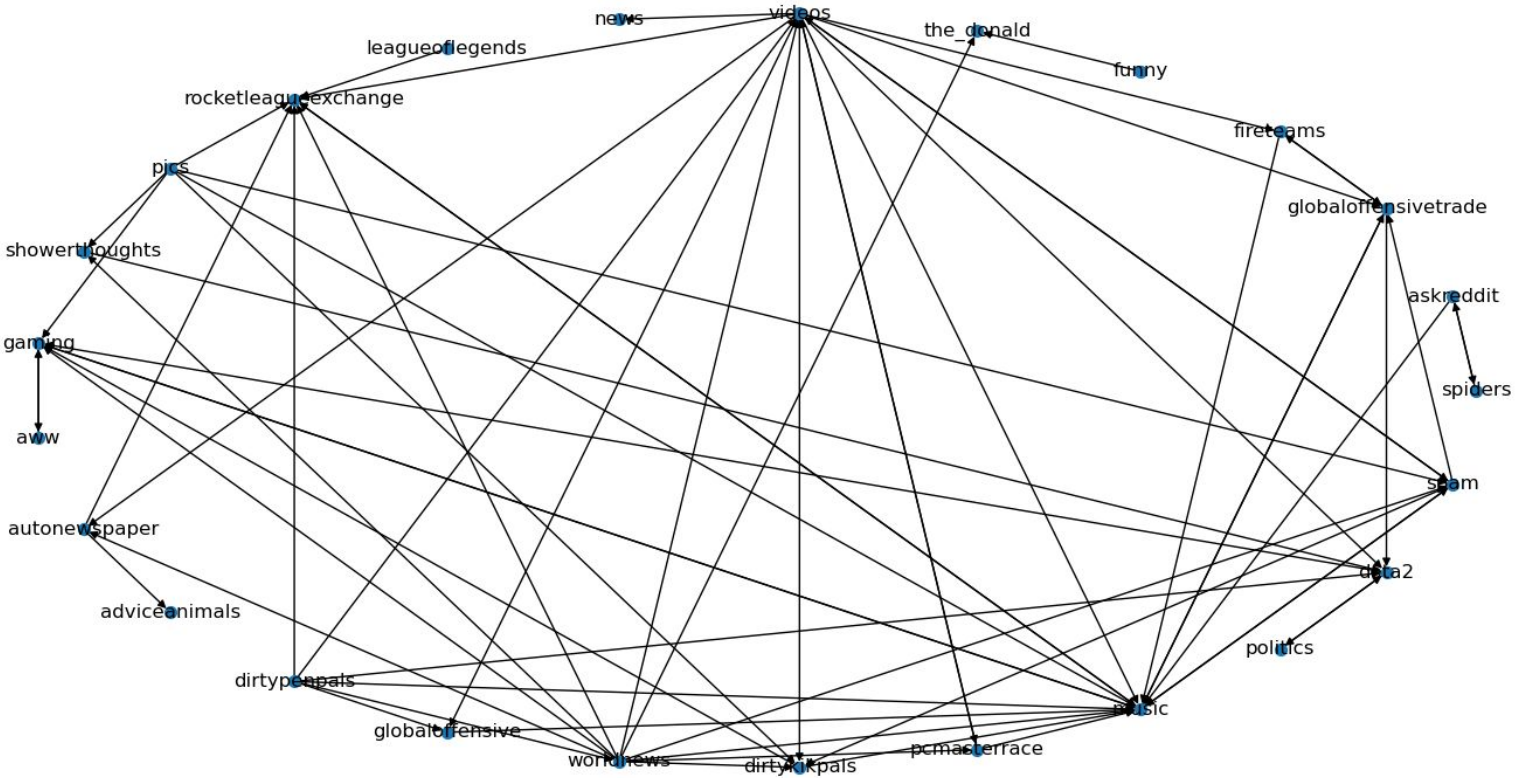
Aayush Sharma (ASU ID: 1222229268)

Rajath Manjunath (ASU ID: 1222210977)

PROBLEM FORMULATION

- **Reddit** is a social news aggregation and discussion website, where members post content like text, posts, images etc.[1]
- Posts are organized by subject boards called **"communities" or "Subreddits"**, which cover various topics such as news, politics, science, movies, video games etc.[1]
- "Subreddits" become a channel of propagation of information when a post at one subreddit is **linked or tagged** in another subreddit.
- We formulate this network as a Graph, where **each node is a "Subreddit"** and **edges are existing hyperlinks** between them. **("Subreddit-Sphere")**
- Given a **pair of nodes** ("Subreddits") in that graph, we wish to predict the **probability** that of one subreddit hyperlinking or tagging a post in the other subreddit.



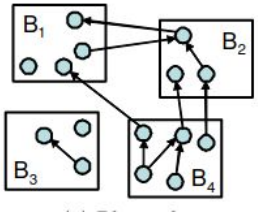


Connection Graph of First 20 Subreddits

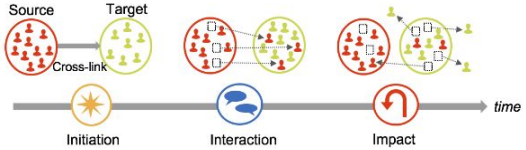
The edge link direction represents the subreddit hyperlinking another subreddit.

PREVIOUS WORKS

Our motivation for working towards this project came from two substantial works:



- Patterns of Cascading Behavior in Large Blog Graphs[3] by Leskovec et al.
  - Formulated a **cascading pattern for the propagation** of information in the blog-to-blog network.
  - The probability of propagation was **constant irrespective** of the node features and links.
  - **Results were obtained from empirical observations.**



- Community Interaction and Conflict on the Web[4] by Kumar et al.
  - Studied intercommunity (subreddits) **interactions and sentiment analysis** of the hyperlinks on the posts.
  - Used an **LSTM Architecture** for the possibility of conflicts between two communities.
  - **Directed towards sentiment and conflict analysis.**

DATASET

- We model our graph based on the **Social Network: Reddit Hyperlink Network** dataset.[2]
- The **subreddit-to-subreddit hyperlink** network is extracted from the posts that create hyperlinks from one subreddit to another.
- The network is **directed, signed, temporal, and attributed.**
- Each **hyperlink** is annotated with three properties:
  - The timestamp
  - The sentiment of the source community post towards the target community post.
  - Text property vector of the source post.
- The dataset contains three sub-datasets:
  - Network of subreddit-to-subreddit hyperlinks extracted from **hyperlinks in the body of the post.**
  - Network of subreddit-to-subreddit hyperlinks extracted from **hyperlinks in the title of the post.**
  - **Subreddit Embeddings:** Embedding vectors representing each subreddit. (300 dimension vector)

DATASET PREPROCESSING

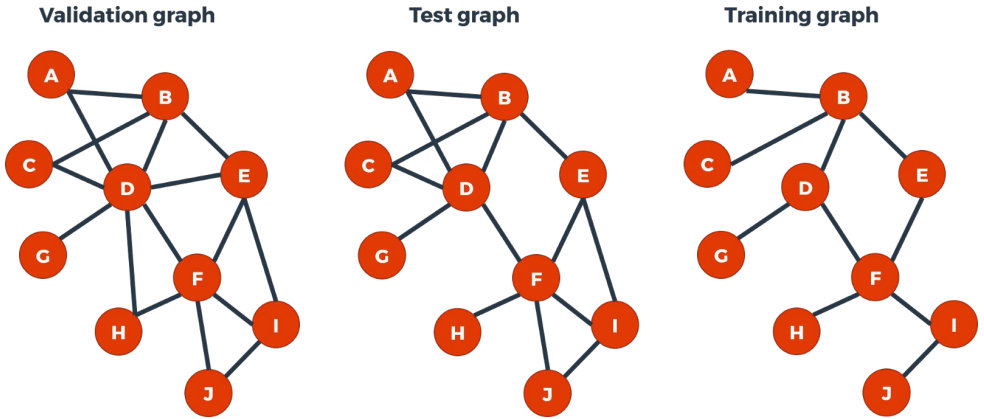
- **Combined** the network of hyperlinks extracted from **hyperlinks in the body of the post** and **hyperlinks extracted from title of the post** into one dataset and keep **only** the **SOURCE** and **TARGET** label attributes.
- **Normalized** the vector **embeddings** of the subreddits.
- We argue that vector embeddings of the subreddits and their past interactions are **enough to predict the probability** of link between two subreddits.

	SOURCE_SUBREDDIT	TARGET_SUBREDDIT
0	rddtgaming	rddtrust
1	xboxone	battlefield_4
2	ps4	battlefield_4
3	fitnesscircle	leangains
4	fitnesscircle	lifeprotips

Sample Links in the Graph

GRAPH CREATION

- We call the original graph as the **Validation Graph.** It is the exact graph formed from the original dataset.
- The graph is modelled using the **NetworkX** library.
- Created a **Test Graph** which is subset of the **Validation Graph**. (Containing 90% of the links from the Validation Graph).
- Created a **Training Graph** which is subset of the **Test Graph**. (Containing 70% of the links from the Test Graph).

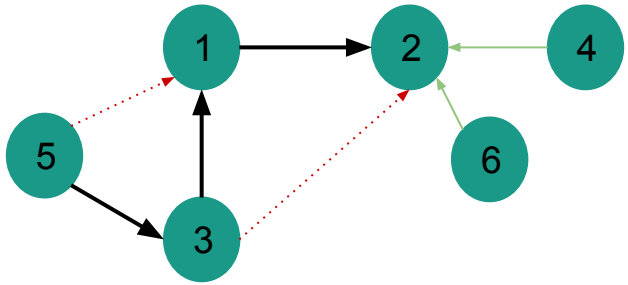


Sample Figure of Validation, Test and Training Graphs

277041 249336 174535

DATASET CREATION

- We consider the base case of link prediction between two subreddits when the **Hopping Distance** between them is **2.**
- We define **Hopping Distance** as shortest path between two nodes other than the direct edge between them (If any).



Hopping Distance between **3-2** is **2.**

Hopping Distance between **5-1** is **2.**

- We use NetworkiT **"MissingLinksFinder"** implementation on both the **Test Graph** and **Training Graph** , to find such pairs.
- We have our **Training and Test Set!!!**
- Please note that these pairs **might or might not have direct links** between them.
- If direct link exists, label the pair as '1' or else label it as '0'.

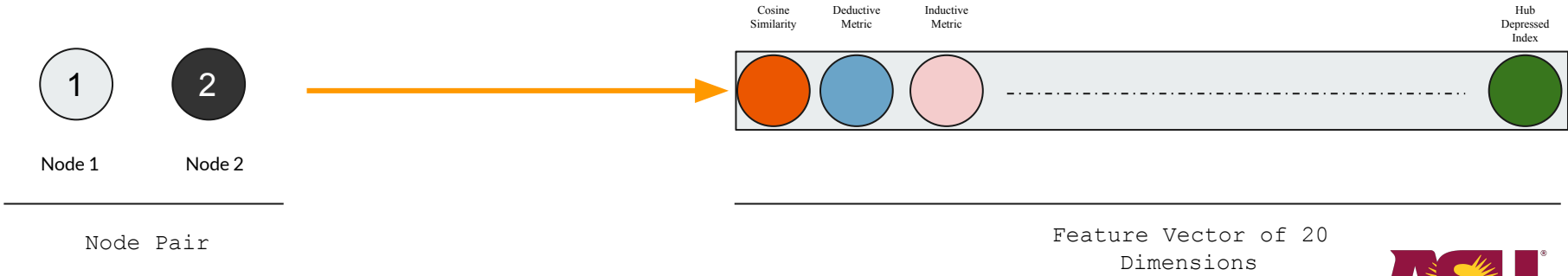
DATASET CREATION

- Consider the Training Set for a moment.
    - **6903 pairs** which have have a label of 1
    - **5359669 pairs** which have a label of 0.

- Classical Case of **Imbalanced Dataset**
- Mitigate this issue by removing pairs of the set with label ‘0’ with a probability of 0.8.
- Following this method for method:
    - **6903 pairs** which have have a label of 1
    - **1077810 pairs** which have a label of 0.
- Data points were not removed from the Test Set and therefore removing additional pairs with label ‘0’ from training set might affect the algorithm’s efficiency to correctly predict True Missing Links and thus leading to some False Positives

FEATURE ENGINEERING

- **Subreddit Embeddings** as node's **vector representation** in the graph.
- Given a pair of nodes (i.e. vector representation) we use various feature engineering techniques, to create feature vector of each pair of nodes.
- We consider **20 pre-defined individual metrics** on each node-pairs.
- These input to each of these metrics are the **vector representations** of each of the node in a pair.
- Concatenating the output from each metric, we get a 20-dimensional vector. The vector essentially becomes the feature vector of the sample.



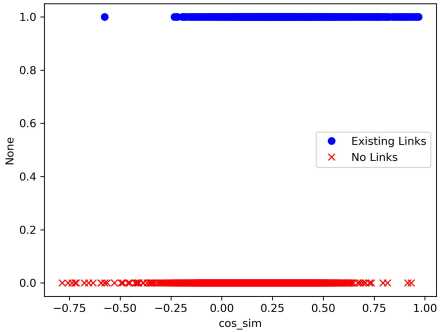
FEATURE ENGINEERING

Deductive Metric[8]	follows deductive reasoning and supported by the generalisations of a node
Inductive Metric[8]	follows inductive reasoning and supported by the specialisations of a node
Deductive_Log metric[8]	Modified Ded metric with a logarithmic term
Inductive_Log metric[8]	Modified Ind metric with a logarithmic term
Inductive+Deductive (INF) Score[8]	Combine both DED and IND into a single score aggregating the evidence provided by ancestors & descendants of a vertex
Ind + Ded with 2xDed Score[8]	modified INF as INF_2D where DED score is given twice the weight of the IND score.
Ind + Ded with 2xINF Score[8]	Modified INF as INF_2I where IND score given twice the weight of the DED score.
INF_LOG Score[8]	A modification of INF taking into account, the satisfying ancestors and descendants.
INF_LOG_2D Score[8]	INF_LOG_2D is INF_LOG score given twice the weight of the IND_LOG score
INF_LOG_2I Score[8]	INF_LOG_2I is INF_LOG score given twice the weight of the DED_LOG score

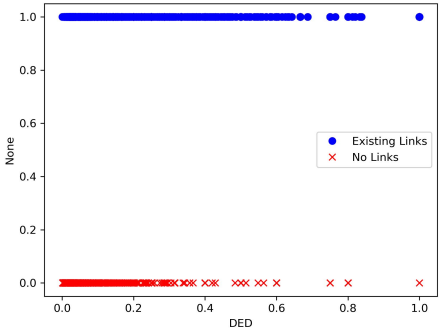
Cosine Similarity[5]	For each pair of nodes with common neighbours, a dot product of vectors is done
Common neighbors score [5][6]	Number of shared nodes between 2 nodes is the similarity between them
Resource allocation index[6]	Inspired from the resource allocation process taking place in complex networks
Jaccard Coefficient[5]	ratio of shared neighbors to the complete set of neighbors for two nodes.
Adamic Adar Index[5]	each feature weight is logarithmically penalized by its appearance frequency.
Preferential Attachment[5]	As degree of the nodes increases, the probability of link formation also increases
Salton Index[7]	The Salton index yields a value that is approximately twice the Jaccard index.
Sørensen Index (SO)[5]	Similar to Jaccard coefficient but less susceptible to outliers
Hub Promoted Index (HPI)[5]	promote link formation between low-degree nodes and hubs.
Hub Depressed Index (HDI)[5]	promotes link formation between hubs and between low-degree nodes

FEATURE ENGINEERING

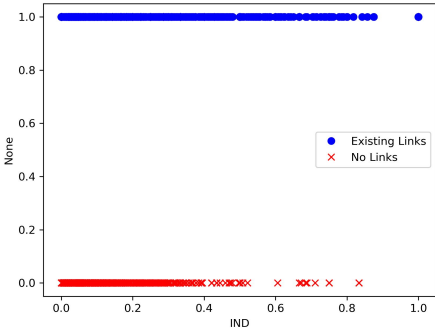
Plotting **Individual Metrics** on the x-Axis



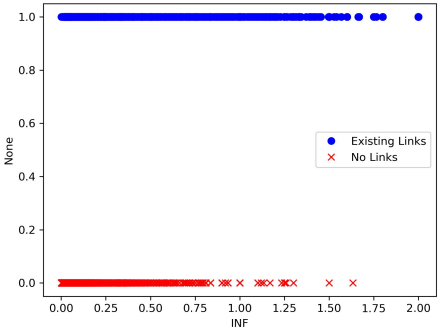
Cosine similarity



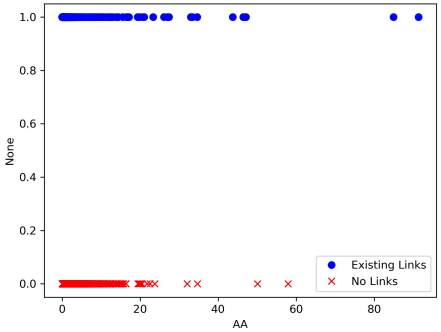
Deductive score



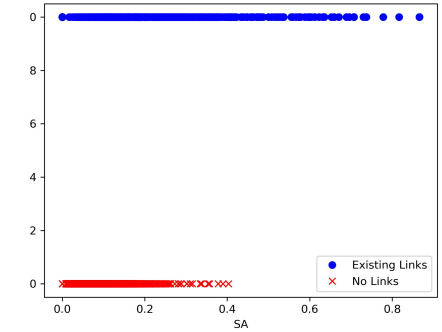
Inductive score



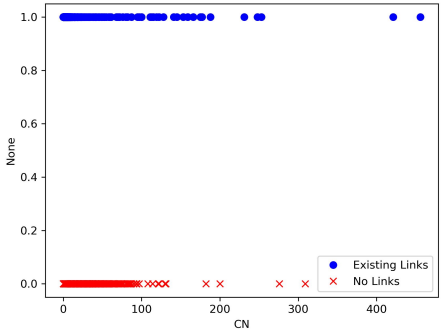
Inductive + deductive score



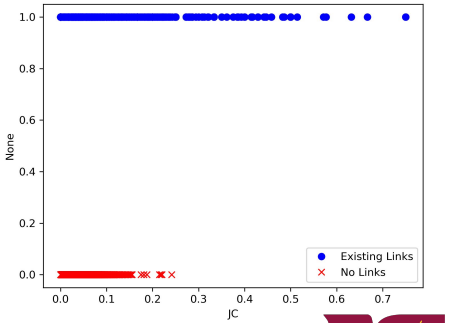
Adamic Adar score



Salton Index



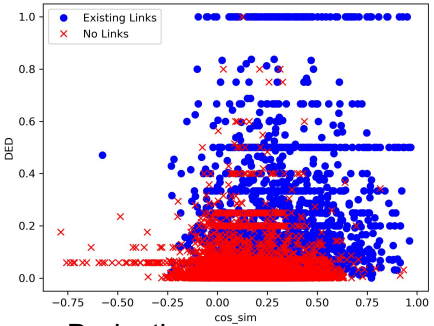
Common neighbors score



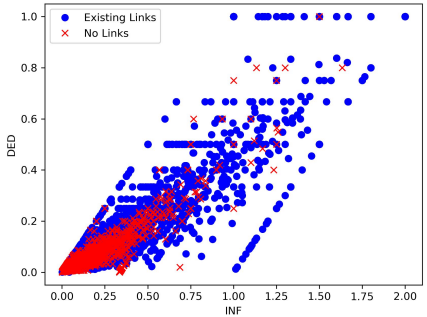
Jaccard Coefficient

FEATURE ENGINEERING

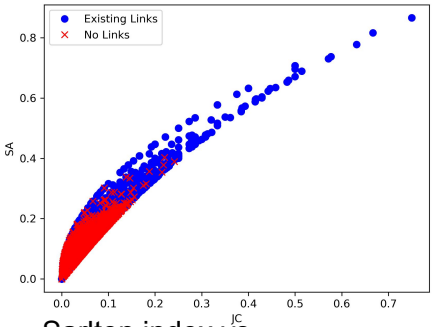
Plotting **Two Metrics** on the x- Axis and the y-Axis



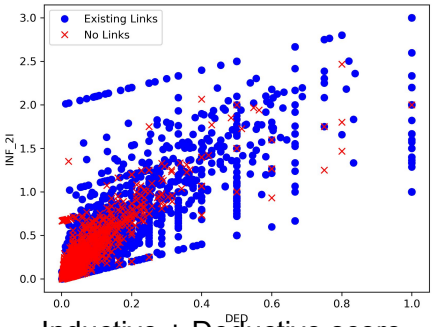
Deductive score vs Cosine similarity



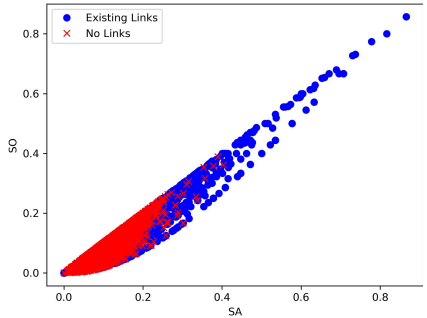
Deductive score vs Inductive score



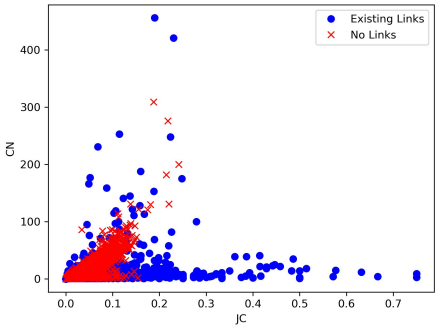
Sarlton index vs Jaccard coefficient



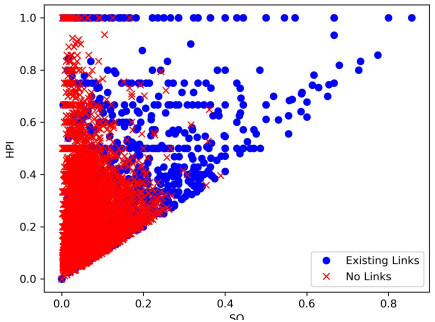
Inductive + Deductive score vs Deductive score



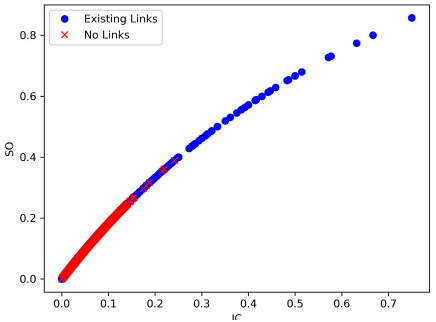
Sørensen Index vs Sarlton Index



Common neighbors score vs Jaccard Coefficient



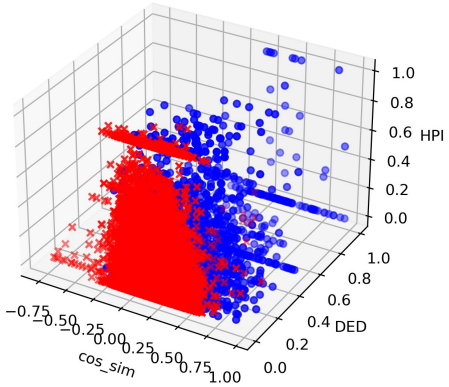
Hub Promoted Index vs Sørensen Index



Sørensen Index vs Jaccard coefficient

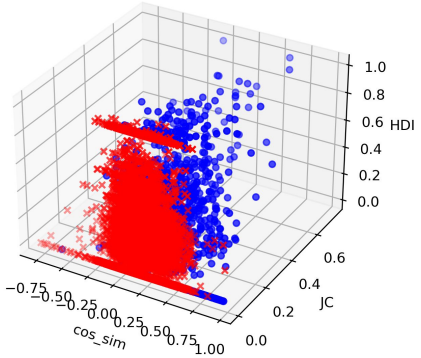
FEATURE ENGINEERING

Plotting **Three Metrics** on the x- Axis, y-Axis and the z-Axis

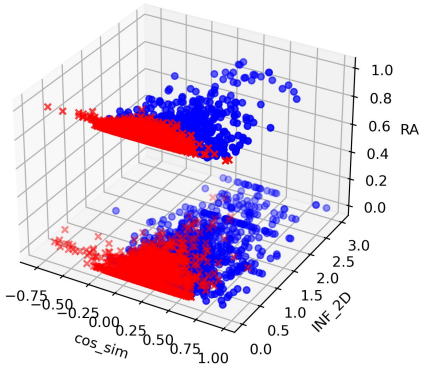


Hub Promoted Index
vs
Deductive Score
vs
Cosine similarity

Cluster being Formed



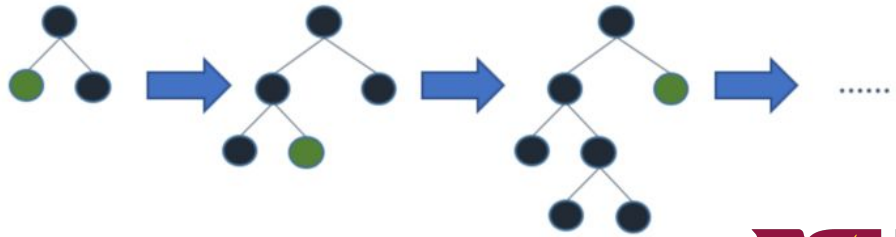
Hub Depressed Index
vs
Jaccard Coefficient
vs
Cosine similarity



Resource Allocation Index
vs
Ind + Ded with 2xDed score
vs
Cosine similarity

LINK PREDICTION ALGORITHM

- Given the feature vector representing a pair of nodes (Subreddits), we wish **predict if there exists a link** between them or shall exist in future.
- Link is defined as the edge direction from the **Source Node to the Target Node.**
- Consider the problem as **binary classification problem (0 or 1)**.
- We use **Light Gradient Boosting Method(LGBM)** for this classification problem.
- Light GBM is a gradient boosting framework that uses tree based learning algorithm.
- LGBM is a preferred tree based learning algorithm for few reasons:
  - **High speed.**
  - **Handle the large size**
  - **Takes lower memory to run**.



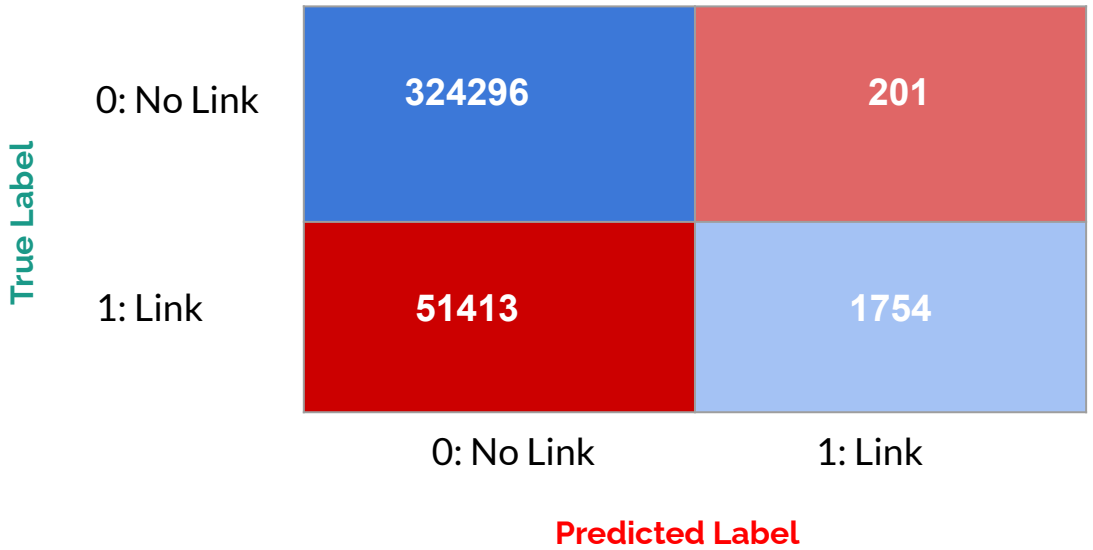
Leaf-wise tree growth

RESULTS

Number of Training Samples: 1077810

Number of Testing Samples: 1453519

Best AUC Score on Test Samples: 0.876927



CONCLUSION AND FUTURE WORK

- We were successfully able to model the problem of link prediction with a **Hopping Distance** of **2**.
- Our main contributions in the work were as follows:

    1) Designed a **Feature Vector Based** on Classical Link Prediction Algorithms.

    2) Used **Gradient Boosting Method** for Link Prediction in the Graphs.

    3) Provided Empirical Results to show good results can be achieved in link prediction problems without the use of more sophisticated node-feature extraction methods such as **node2vec.**[9]

- For future work we wish to experiment with higher **Hopping Distance like 3,4 5** etc. and evaluate the algorithm's effectiveness in such cases.
- With greater hopping distances many of the individual features **MAY NOT** be effective enough to model the network to predict the link between the nodes.

REFERENCES

[1] Wikipedia https://en.wikipedia.org/wiki/Reddit

[2] Social Network: Reddit Hyperlink Network https://snap.stanford.edu/data/soc-RedditHyperlinks.html

[3] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading Behavior in Large Blog Graphs. Proceedings of the 2007 SIAM International Conference on Data Mining (SDM). 2007, 551-556

[4] S. Kumar, W.L. Hamilton, J. Leskovec, D. Jurafsky. Community Interaction and Conflict on the Web. World Wide Web Conference 2018.

[5] Link Prediction Methods and Their Accuracy for Different Social Networks and Network Metrics /www.hindawi.com/journals/sp/2015/172879/

[6] Comparing Sets of Pattern with Jaccard Index https://www.researchgate.net/publication/323624308

[7] Similarity Measure for Link Prediction in Social Network https://www.researchgate.net/publication/339754068

[8] Link Prediction in Very Large Directed Graphs http://ceur-ws.org/Vol-1243/paper5.pdf

[9] node2vec: Scalable Feature Learning for Networks. A. Grover, J. Leskovec. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.