Political Alliance Prediction using Twitter Data

-By Data Diggers

Abstract

We study the problem of predicting the political alliance of parties on the Twitter network, showing that the political alliance of parties can be predicted from their Twitter behavior. The work involves applying computational politics over a large dataset including Tweets and mention network of Indian Politicians. In this study, we have taken both structural and content similarity to predict the alliance. First, we analysed the networks using a tool called Gephi. Secondly, We calculated the EI index of inter communities, content similarity and sentiment of tweets. Finally, we combined the results of those measurements and predicted the probability of the alliance. According to our findings, we predict that there is a high probability of alliance between two political parties which are INC-AAP, INC-RJD.

Keywords: Political Alliance Prediction, Network Analysis, EI index, Cosine Similarity, Jaccard Similarity.

1. Introduction

The twitter data has entities and attributes in its datasets which mainly includes a user and its tweet. A tweet has several features like Textual feature, Topic Feature, non textual feature and thread features. On the other hand the user has social features and non social features. To measure and analyse all these features is the task of this work.

Computational politics means to apply computational methods over a large dataset obtained from offline or online sources for conducting outreach, mobilisation in service of electing , supporting or opposing a party, policy or legislation [1]. The methods in computational methods include statistical analysis, probabilistic models and behavior of users. These methods help measure the socio-political behavior and ideologies between different communities. These insights further help for opinion mining, polls, marketing and prediction of coalitions. Computation politics is the umbrella which carries tasks such as Community and user modelling, Information flow, political discourse, Election campaigns and system design. User modelling is often defined by keywords that include homophily, political affiliation, influencers and many more to define the users of a community. Information flow contains misinformation like bias, fake and propaganda as major pillars to define the misconduct of information. This deals with data propagation and information flow in the media. Other tasks often involve opinion mining and topic modelling.

1.1 Twitter Data and its features : Users share messages called tweets within their network. These tweets have four attributes. Textual features include length,sentiments and writing style and the message conveyed. Non-textual features are composed of the tweets

metadata. Thread features highlight the information flow in the network. Whether a tweet is a mention or retweet. The topic features restrictions over length of tweet, the hashtags of tweets. In our work we analysed the twitter data over three hashtags named COVID, Farmbills and IndoChina.

1.2 Community and User Modelling : The online behavior of a user as an individual and as a community can be detected by community and user modelling. It deals with propagation of information within the network. The term ***Homophily*** refers to the phenomenon where people with similar interests form a connection with each other. The ideology of the community is similar. It follows the proverb "birds of a feather flock together". ***Political affiliations*** means the behavior of users where they tend to show positive sentiment towards some political entity. ***Influncers*** in social network analysis are the users whose content has the largest reach and are mostly the ones whose actions and thoughts can affect the behavior of other people.

1.3 Information flow : This approach deals with the flow of information and its propagation within and outside a network. Here, ***Misinformation*** specialises in detection of bias and fake news along with some fact checking. ***Echo chamber***, refers to a situation where the beliefs of a community are amplified by repetition and communication within a closed community.

1.4 **Network analysis :** The network analysis can be done based on various centrality measures, which tells us how important a node or a user is in the social network analysis. The popular centrality measures are Degree centrality, closeness centrality, betweenness centrality, eigenvector centrality and pagerank centrality. We used some of these to analyse our twitter data. **Degree centrality** is the measure of node connectivity, which measures this based on in-degree and out degree of the node. The nodes with higher degree are considered to be central. This measure is used to find connected individuals, popular individuals. **Betweenness Centrality** is defined for a nodes as the number of times a node lies on the shortest path between other nodes. This is used to find the individual who can influence the flow in the system.
Modularity is one measure of the structure of networks or graphs which measures the strength of division of a network into modules (also called groups, clusters or communities). Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules.

2. Methodology

For Network Analysis we used a tool called GEPHI. For analyzing the graphs we divided the dataset into edgelist and nodelist. The graphs were visualised over a layout called Force Atlas 2. The communities were detected based on modularity. Each community was given a modularity class. The graphs were then analysed on different centrality measures. The dataset was analysed and graphs were plotted against politicians and parties. We considered three different cases for plotting the graphs based on the users. Which were Politician v/s politician, Politician v/s party and party v/s Party.

Network analysis graphs were visualised over two different types of network for each of the above constraints : retweet and mentions networks. For centrality measures we considered betweenness centrality and degree centrality. On the basis of these visualisations we formed our results for the most influential person in a network. Also, depiction of endorsement networks was observed from the retweet networks. The *mentions* network was used for analysing the sentiment of the community.

Our method is based on the following three factors:

a) EI index of inter communities
b) Content Similarity of two communities
c) Sentiments of the tweets in a particular topic *(Covid/FarmBill/IndoChina)*

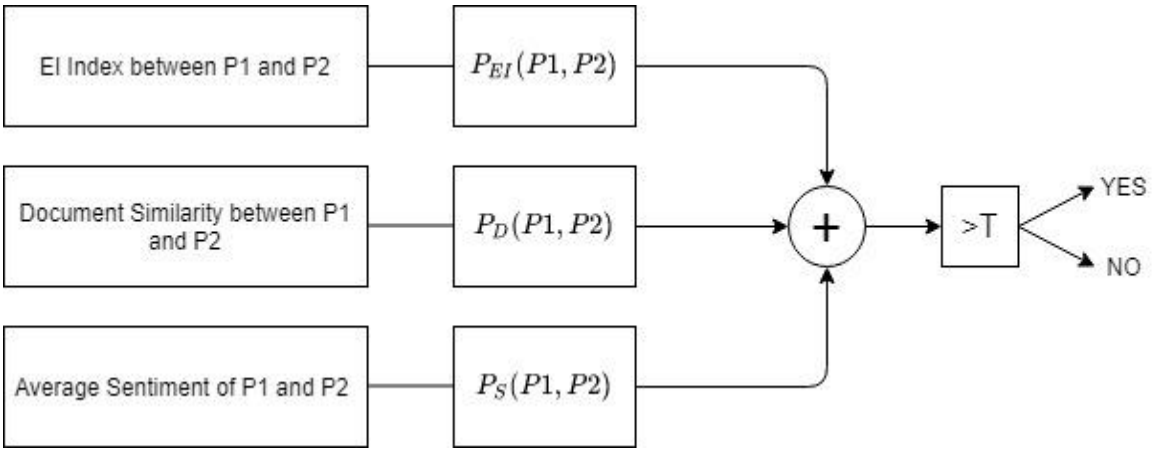

Fig 1: Idea for proposed implementation

2.1) EI index of inter communities: The EI index between two communities have been calculated by considering the number of internal and external edges they have.
We have used the formula EI = (E-I)/(E+I) to calculate the EI index. Where E is the number of external edges and I is the number of internal edges of a party, among a group of parties.

2.2) Content Similarity of two communities: We have used *cosine* and *jaccard* similarity to measure the similarity between two documents.

2.2.1) *Cosine Similarity* for documents A and B of two different parties is given by:

$$Sim(A, B) = Cos(\theta) = A.B \div |A||B|$$

2.2.2*) Jaccard Similarity* for documents A and B of two different parties is given by:

$$Sim(A, B) = |A \cap B| \div |A \cup B|$$

2.2.3) *Euclidean Distance*: $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

2.3.4) *Manhattan Distance*: $|x_1 - x_2| + |y_1 - y_2|$

3. Implementation Details and Result

For network analysis we used Gephi. We used Python programming and also used Microsoft excel for filtering and formulating our dataset. Different python libraries used were nltk, sklearn, scipy, numpy, word2vec, textblob. We have only considered those parties whose tweets are more, few parties are present with very few tweets, also those pirates are not present in all the topics (covid, farmbill, indo-china) .

3.1. Network Analysis:

- Based on Hashtags the dataset was divided into three subsets. For Covid the number of hashtags used were 46, for FarmBill the number of hashtags used were 68 and for IndoChina it was 28.
- From the politician.txt we first separated out the politicians and parties
- Mapped Politician to its respective Party (945)
- Filtered the dataset based on the given indian politicians and pirates
- Prepared Edgelist and Nodelist for Gephi and visualized graphs
- Used the above steps for community detection.

3.2 Data Statistics:

- Topics : COVID, FARMBILL, INDOCHINA
- Number of tweets in each Topic
    1. Covid : 11300
    2. Farmbill : 2107
    3. IndoChina : 125
- Data for each topic according to the network type in Politician v/s Politician

Topic (Pol v/s Pol)	Number Of Tweets	Number of Politicians	Number of Parties
Covid Mentions	4785	519	30
Covid Retweet	1386	364	24
FarmBill Mentions	1038	173	16
FarmBill Retweet	288	75	7
IndoChina Mentions	51	35	6
IndoChina Retweet	10	9	3

3.3. Visualisations:

The Visualisation results for each Network from Gephi

3.3.1). COVID Mention Network

	Nodes, Edges	Modularity	Graph Density
Politician vs Politician	524,1370	0.555	0.005
Politician vs Party	143,152	0.798	0.007
Party vs Party	23,22	0.545	0.043

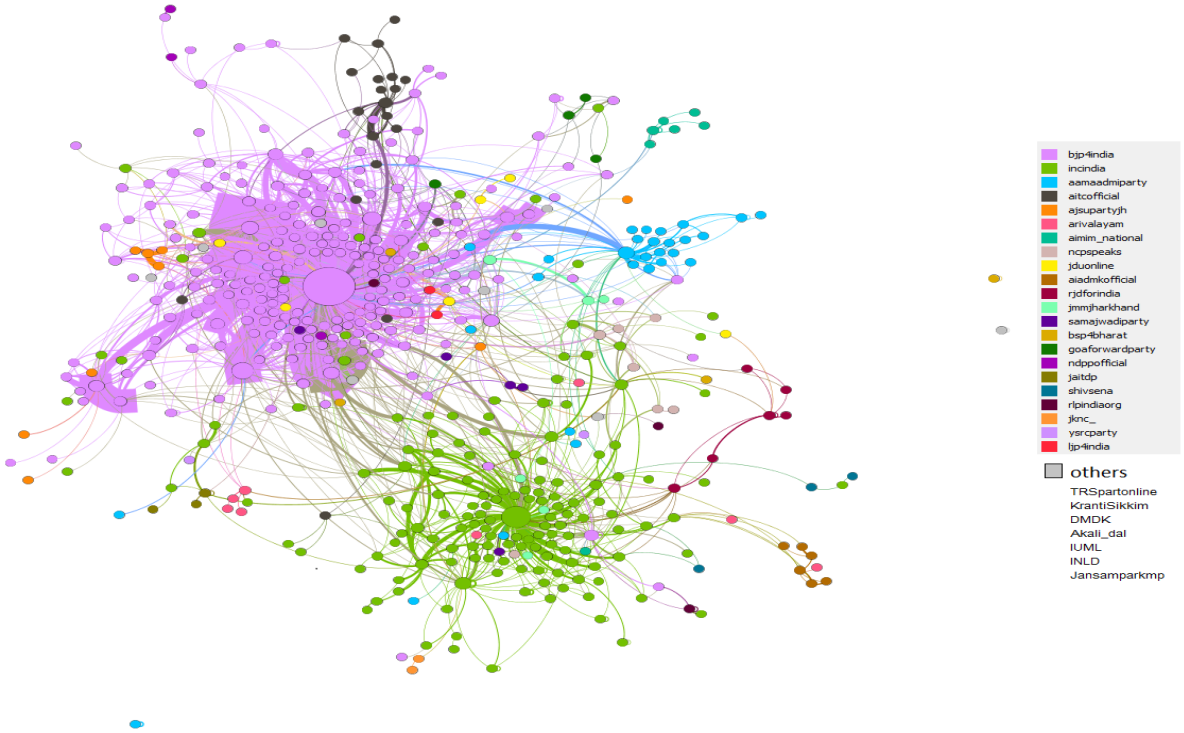

Fig. 2: Covid mention politician vs politician

3.3.2) COVID Retweet Network

	Nodes, Edges	Modularity	Graph Density
Politician vs Politician	359,576	0.683	0.004
Politician vs Party	98,88	0.774	0.009
Party vs Party	15,13	0.392	0.062

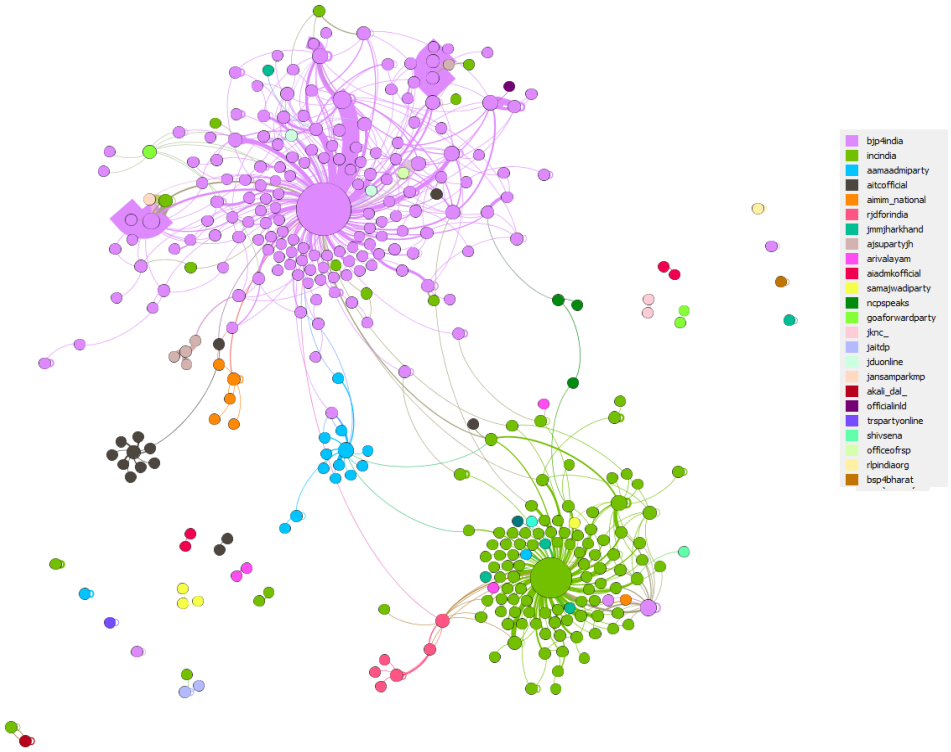

Fig. 3: Covid Retweet politician vs politician

3.3.3). FarmBill Mentions Network

	Nodes, Edges	Modularity	Graph Density
Politician vs Politician	173, 689	0.213	0.023
Politician vs Party	33,33	0.777	0.031
Party vs Party	2,3	0	1.5

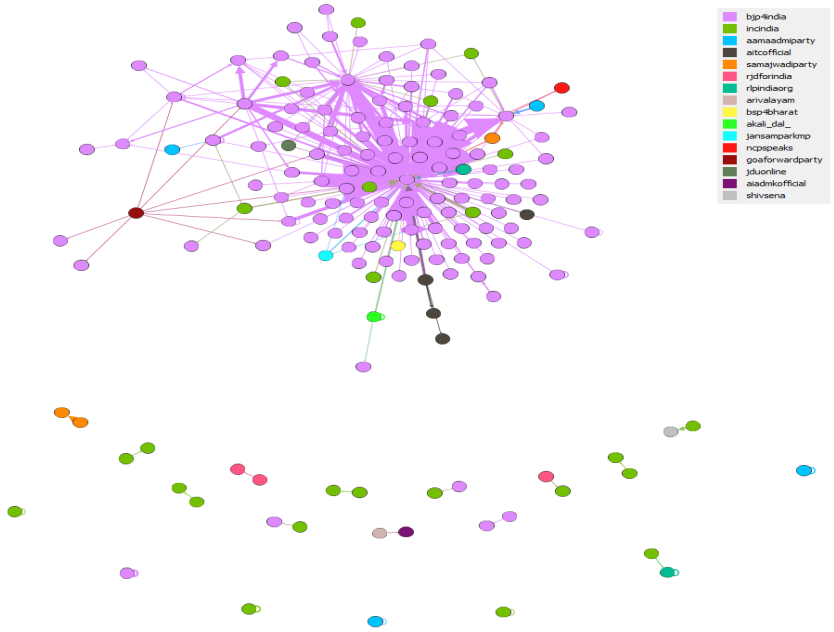

Fig. 4: FarmBill mention politician vs politician

3.3.4) FarmBill Retweet Network

	Nodes, Edges	Modularity	Graph Density
Politician vs Politician	75,107	0.607	0.019
Politician vs Party	13,8	0.750	0.051
Party vs Party	2,3	0	1.5

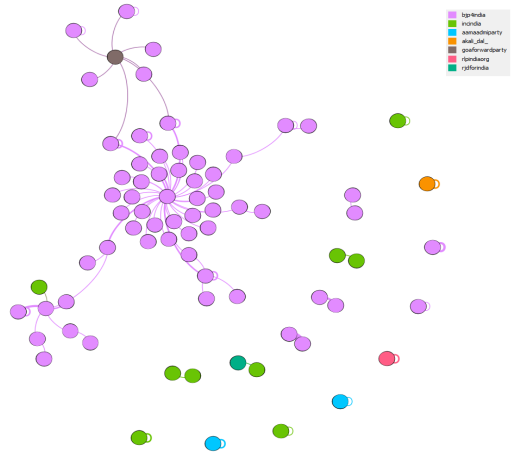

Fig. 5: FarmBill retweet politician vs politician

3.3.5). Indochina Mentions Network

	Nodes, Edges	Modularity	Graph Density
Politician vs Politician	35,51	0.585	0.043
Politician vs Party	10,10	0.66	0.111
Party vs Party	4,4	0.444	0.333

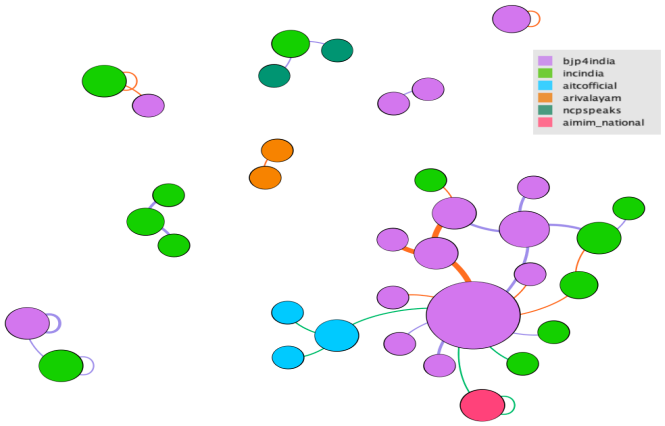

Fig. 6 : IndoChina Mentions Politician vs Politician

3.3.6). Indochina Retweet Network

	Nodes, Edges	Modularity	Graph Density
Politician vs Politician	9,10	0.75	0.139
Politician vs Party	2,1	0.0	0.5
Party vs Party	2,2	0	1

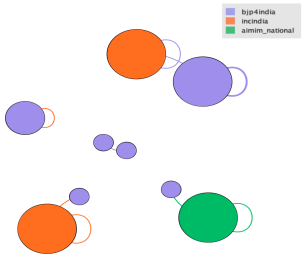

Fig. 7: IndoChina Retweet Politician vs Politician

3.4 EI INDEX

External internal Index is a social network measure which measures the relative density of internal connections within a social group compared to the number of connections that group has to the external world. We calculated the EI index within the communities and across the communities.

3.4.1) EI index of the entire Dataset

Covid - Mentions	-0.5842441740565065
Covid - Retweet	-0.6236481614996395
FarmBill - Mentions	-0.8026124818577649
FarmBill - Retweet	-0.8317757009345794
IndoChina - Mentions	-0.5294117647058824
IndoChina - Retweet	-0.4

3.4.2) EI Index for COVID Within Communities

	Mentions	Retweet
BJP	-0.8964	-0.9249
AAP	-0.4328	-0.5483
INC	-0.0799	-0.6463
AIADMK	-1.0	-1.0
AITC	-0.2580	-0.625
arivalayam	0.5789	-0.33
Akali dal	1	1

3.4.3) EI Index for FarmBills Within Communities

	Mentions	Retweet
BJP	-0.9933	-1.0
INC	0.5294	-0.7142
AAP	0.1428	-1.0
AITC	0.25	0
Akali Dal	0.2	-1.0

3.4.4) EI Index for IndoChina Within Communities

	Mentions	Retweet
AAP	0	0
BJP	-0.8620	-1.0
INC	-0.1428	0
AITC	-0.3333	0
AIADMK	0	0

3.4.5) EI Index Inter Community for mention networks

	Covid	FarmBill	IndoChina	Average EI Index
BJP - AAP	-0.9810	-1.0	-1.0	-0.9936
BJP - AITC Official	-0.9854	-1.0	-1.0	-0.9951
INC - AAP	-0.51278	0.12923	0.23	-0.0511
INC - AITC Official	-0.32915	-0.23977	-1.0	-0.5229

INC - RJD	-0.18105	0.22983	0.0	0.02439
BJP - RJD	-0.9974	-0.9933	0.0	-0.99535
AAP - AITC Official	-0.69038	-0.39	-1.0	-0.6934

Inferences: It is observed that for Covid and FarmBill the *retweet* network is more polarized whereas for IndoChina *mentions* network is more polarized.
It is also observed that BJP is having highly negative average EI Indexes. Hence we can say that BJP is more polarized.

3.5 Influential Community

3.5.1). COVID Mention Politician vs Politician:

- Based on In-degree: Narendra Modi (235), Rahul Gandhi (109)
- Based on Betweenness: Dr. Harsh vardhan (13012.38), ML Khattar(3004.083)

3.5.2). COVID Retweet Politician vs Politician:

- Based on In-degree: Narendra Modi (129), Rahul Gandhi(95)
- Based on Betweenness: Finance Minister India(76.5),Chouhan Shivraj(58.5)

3.5.3). FarmBill Mention Politician vs Politician:

- Based on In-degree: Narendra Modi (405), nstomar (86)
- Based on Betweenness: Narendra Modi (82.59), nstomar (7.43)

3.5.4). FarmBill Retweet Politician vs Politician:

- Based on In-degree: Narendra Modi (38), nstomar (9)
- Based on Betweenness: piyushgoyal (3.0), pcmohonmp(2.0)

3.5.5). IndoChina Mention Politician vs Politician:

- Based on In-degree: narendra modi (16), amit shah(7)
- Based on Betweenness: N/A (all zero)

3.5.6). IndoChina Retweet Politician vs Politician:

- Based on In-degree: gvlnrao (3), asadowaisi (1)
- Based on Betweenness: N/A (all zero)

3.6 Sentiment Analysis

From the dataset we were given three sentiments majorly 0 (+ve), 1(-ve), 2(N). Also we have used TextBlob for sentiment analysis. TextBlob aims to provide access to common text-processing operations through a familiar interface. The sentiment property returns a namedtuple of the form Sentiment(polarity, subjectivity). The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective. We consider only the polarity score for our analysis.

	Covid	FarmBill	IndoChina	Average Prob.
BJP - AAP	0.8292	0.9537	0.7551	0.8460
BJP - AITC Official	0.7734	0.9300	0.2900	0.6644
INC - AAP	0.9085	0.8861	0.8344	0.8763
INC - AITC Official	0.8527	0.9932	0.3750	0.7403
INC - RJD	0.9286	0.9932	0.5000	0.8072
BJP - RJD	0.8493	0.9392	0.5000	0.7628
AAP - AITC Official	0.9442	0.8929	1.0000	0.9457

3.7 Similarity Measures

3.7.1) Similarity and dissimilarity Measures for COVID Dataset:

COVID	Cosine Similarity	Jaccard Similarity	Euclidean Dist.	Manhattan Dist.
BJP - AAP	0.21829	0.066731	102.78	10564.0
BJP - GoaForwardParty	0.16946	0.039414	103.47	10707.0
BJP - AITC Official	0.21980	0.067439	102.69	10547.0

BJP - RJD0	0.14892	0.032153	103.84	10784.0
AAP - AITC Official	0.25737	0.147661	40.33	1627.0
INC - AITC Official	0.25361	0.095835	81.04	6569.0
INC - RJD	0.17958	0.048607	81.87	6704.0
INC - AAP	0.25236	0.095030	81.12	6582.0
INC - GoaForwardParty	0.19778	0.057381	81.62	6663.0

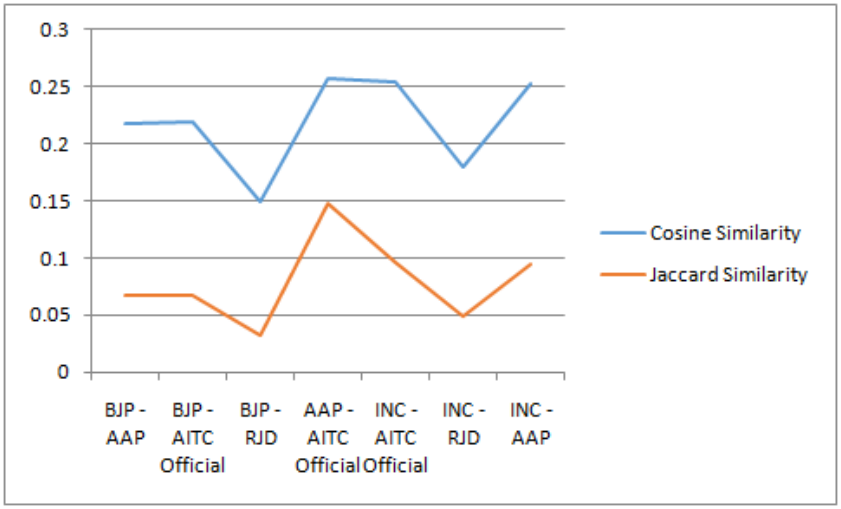

Fig 8: Similarity Measures For Covid Dataset

3.7.2). Similarity and dissimilarity Measures for FarmBill Dataset:

FarmBill	Cosine Similarity	Jaccard Similarity	Euclidean Dist.	Manhattan Dist.
BJP - AAP	0.1589	0.0677	41.88	1754.0
BJP - GoaForwardParty	0.1643	0.0419	39.73	1579.0
BJP - AITC Official	0.0933	0.0171	40.08	1607.0
BJP - RJD	0.1039	0.0268	40.42	1634.0

AAP - AITC Official	0.1404	0.0483	20.32	413.0
INC - AITC Official	0.1282	0.0289	32.21	1038.0
INC - RJD	0.1287	0.0403	32.76	1071.0
INC - AAP	0.2293	0.1142	33.86	1147.0
INC - GoaForwardParty	0.1534	0.0471	32.40	1050.0

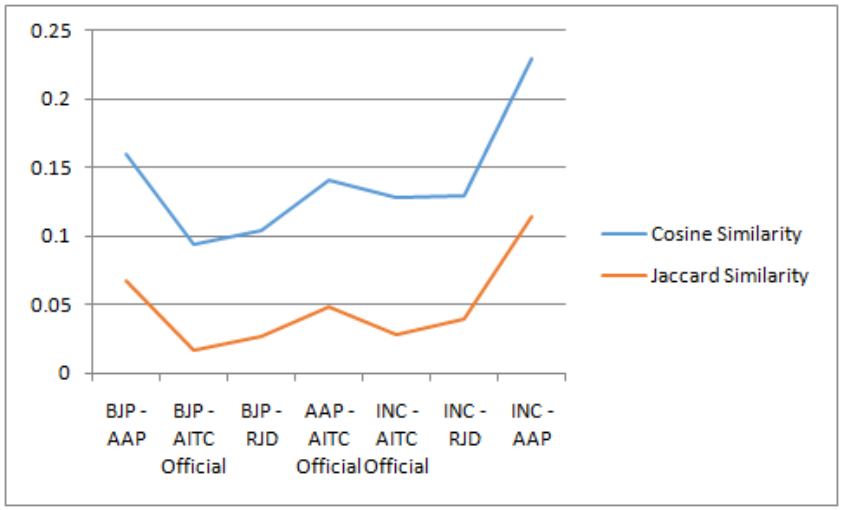

Fig 9: Similarity Measures For Farmbill Dataset

3.7.3). Similarity and dissimilarity Measures for IndoChina Dataset:

IndoChina	Cosine Similarity	Jaccard Similarity	Euclidean Dist.	Manhattan Dist.
BJP - AAP	0.169220	0.045936	23.23	540.0
BJP - GoaForwardParty	0.162180	0.034234	23.15	536.0
BJP - AITC Official	0.177555	0.072463	24.0	576.0
BJP - AIMIM	0.167007	0.058922	23.64	559.0
AAP - AITC Official	0.183280	0.088435	11.57	134.0
INC - AITC Official	0.173561	0.061556	28.42	808.0

INC - AIMIM	0.142570	0.042806	28.37	805.0
INC - AAP	0.118839	0.026894	28.21	796.0
INC - GoaForwardParty	0.134602	0.023661	28.0	784.0

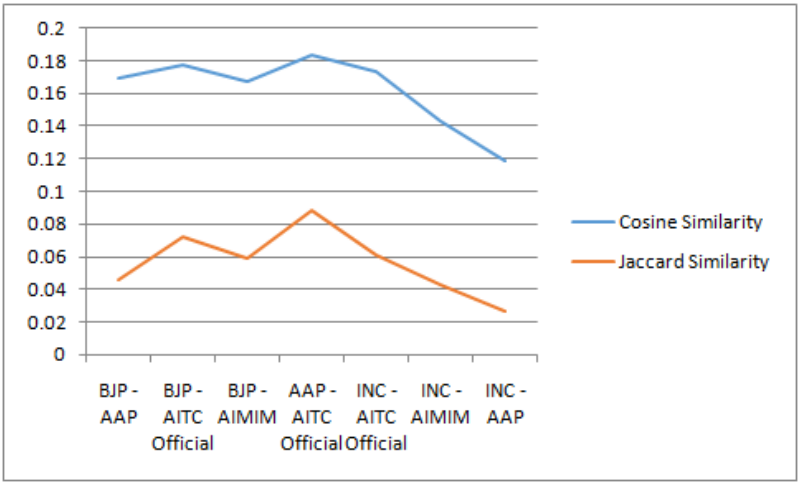

Fig 10: Similarity Measures For IndoChina Dataset

3.8. Probability of Alliance

	Avg. EI Index	Avg. Similarity	Avg. Sentiment	Probability of Alliance
BJP - AAP	0.0030	0.1821	0.8460	0.3437
BJP - AITC Official	0.0020	0.1635	0.6644	0.2766
INC - AAP	0.4744	0.2001	0.8763	0.5169
INC - AITC Official	0.2385	0.1851	0.7403	0.3879
INC - RJD	0.5121	0.1541	0.8072	0.4911
BJP - RJD	0.0023	0.1264	0.7628	0.2971
AAP - AITC Official	0.1533	0.1937	0.9457	0.4309

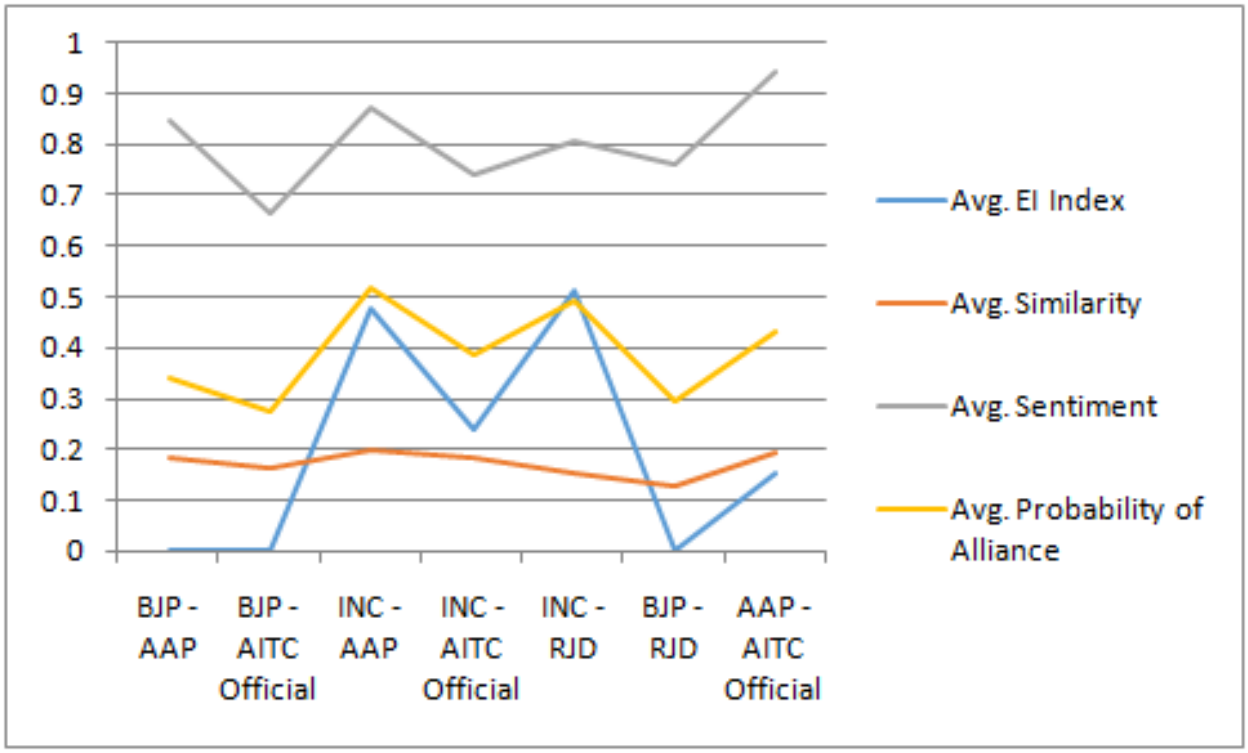

Fig 11: Alliance predictions

Inference: According to our findings, we predict that there is high probability of alliance between two political parties which are INC-AAP, INC-RJD. Also, we observed that most of the parties are supporting INC rather than BJP.

4. Conclusion

From the network analysis we have observed that *retweet* networks are more polarized than the *mentions* networks. Multiple factors are responsible for alliance prediction, we used Structural and Content Analysis to reach our goal, hence used the EI Index, Cosine Similarity and Average Sentiment to calculate the probability of alliance between two political parties and we have reached to the conclusion that INC-AAP and INC-RJD have 51% and 49% probability respectively of forming an alliance. Also from the dataset we observed that some political parties have tweeted much less than others, so we have considered only those political parties that have tweeted a substantial number of tweets. Because the difference between the number of tweets of some parties is quite high, they inherently become less Cosine Similar, hence their probability of alliance becomes low.

5. References

1. https://arxiv.org/abs/1908.06069
2. https://textblob.readthedocs.io/en/dev/quickstart.html
3. https://en.wikipedia.org/wiki/Cosine_similarity
4. https://en.wikipedia.org/wiki/Jaccard_index
5. https://en.wikipedia.org/wiki/Euclidean_distance#:~:text=In%20mathematics%2C%20the%20Euclidean%20distance,being%20called%20the%20Pythagorean%20distance.
6. https://www.sciencedirect.com/topics/mathematics/manhattan-distance
7. https://www.sciencedirect.com/topics/earth-and-planetary-sciences/modularity#:~:text=Modularity%20is%20a%20system%20property,themselves%20rather%20than%20other%20communities.
8. https://faculty.ucr.edu/~hanneman/nettext/C8_Embedding.html
9. https://gephi.org/