1.1 basic analysis

Firstly, the basic features of the four categories of networks in verified Facebook sites are shown as follows, with results round to four decimals:

Tabell 1: basic features of four categories

category	nodes	edges	average_cc	degree_correlation
tvshow	3327	12991	0.3960	0.6321
government	6880	81367	0.4445	0.0240
company	6495	20145	0.2796	0.0285
politician	5768	36909	0.4118	-0.0054

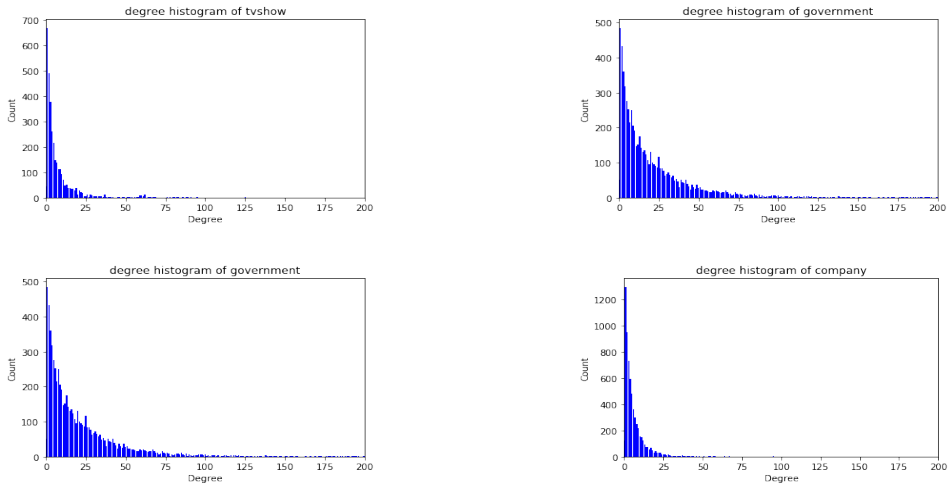
Nodes represent official Facebook pages while the links are mutual "likes" between sites. We can get an intuitive idea of what each category looks like from the table. The government category has the biggest bunch of nodes and edges with relatively high average clustering coefficient, which implies that it's a highly clustered network. By comparison, the company network is much scattered and the tvhsow network is the smallest. Besides, The tvshow category is an assortative network with the highest degree correlation, suggesting that nodes of comparable degree in this category tend to link to each other. Meanwhile the other three categories can be regarded as neutral network and nodes link to each other randomly.

Tabell 2: top three nodes with highest degree and cc

category	node	degree	node	cc
tvshow	4296	125	8208	1.0
	7919	95	16416	1.0
	20516	91	8255	1.0
government	16895	669	9	1.0
	14497	639	10	1.0
	19743	603	15	1.0
company	701	187	63	1.0
	2597	95	125	1.0
	17392	95	198	1.0
politician	14650	323	33	1.0
	20415	253	62	1.0
	21491	222	93	1.0

The top three nodes with highest degree and clustering coefficient are listed on the left. Here we see that the top three nodes in government network have highest overall degree, namely U.S. Army, U.S. Army Chaplain Corps and The White House. The clustering coefficient of several nodes in each category reach 1.0, which means their neighbours form a complete graph. I only pick three in random as an example. We can also notice that the degree of the node is not positive correlated to its clustering coefficient.

The degree distributions of the four categories of network follow power-law, indicating that they are scale-free networks. It can be easily understood: similar to WWW, it's common that a randomly chosen Facebook page has only one or two "likes" from other pages. Yet, it could also be a hub with hundreds of "likes" such as the three nodes mentioned above.



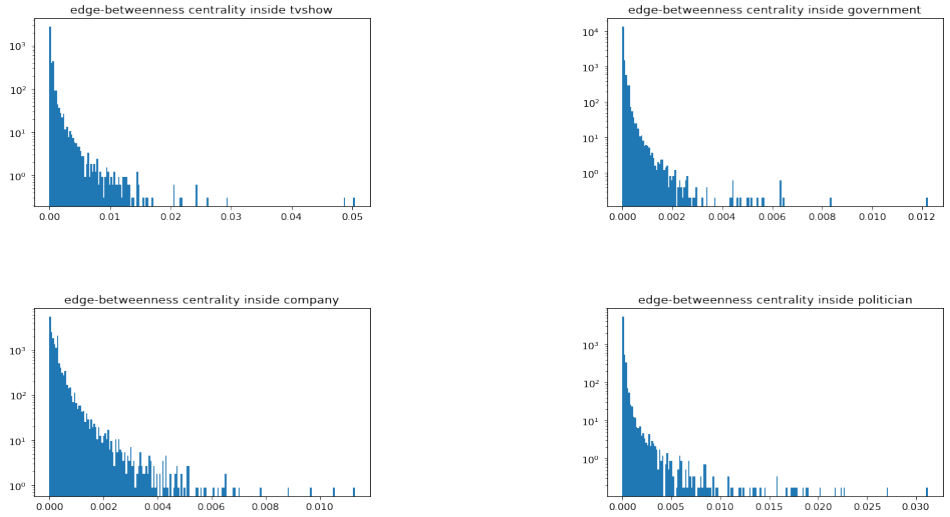
Figur 1: Comparison of degree distribution

1.2 inside and inter-category features

Tabell 3: basic features of four categories

	company	government	politician	tvshow
company	0.8366	0.0265	0.0080	0.0803
government	0.0977	0.9148	0.1088	0.0390
politician	0.0141	0.0520	0.8684	0.0410
tvshow	0.0516	0.0068	0.0149	0.8397

The table shows the fraction of edges from each category going to each other category. We can see that most of the nodes link to other nodes in the same category, especially for the government network. By comparison, The company and politician network have relatively more inter-category edges between each other, the fraction is around 5% and 8% respectively. And the politician network has the highest inter-category edges fraction(10.88%).



Figur 2: Comparison of edge-betweenness centrality

By comparing the edge-betweenness centrality capturing the role of each link in information transfer, we see that the four of them also follow power-law. Only a few hubs are important in information transfer, which means that it would be of great value to get “likes” from them.

Tabell 4: top three nodes with highest inter-category degrees

	tvshow			government			company			politician		
node	909	15839	13425	16052	21120	13098	11332	701	17983	11003	11158	15612
degree	104	58	49	265	187	121	307	193	136	145	95	86

The top three nodes with higest inter-category degrees are shown in 4. It is interesting that NASA in the company category has the highest number of inter-category edges, followed by Senate of Canada and Facebook. Although tvshow category has a relatively large fraction of edges linking to other category, the degree of them are lower than other categories. In contrast, nodes in company and government network have relatively more links in other categories.

2.1 node classifier

With the data I get from part one, I choose three features which contain information for nodes to train the classifier, namely degree, clustering coefficient and centrality. Additionally, I add adjacency matrix as another feature, because the category of the neighbors also implies the category of the chosen node itself. Principal component analysis shows that indeed if connected to certain nodes, it's highly likely to be in the same category. Combining them together, I use train_test_split method to split the combined matrix into train and test sets. I tried three classification algorithms: SVM, Softmax and logistic regression applied on OneVsRestClassifier. The fist one takes hours to compile so I mainly consider the latter two. The logistic regression applied on OneVsRestClassifier gets better results compared to Softmax algorithm. It reaches accuracy of 0.62. And the precision score is relatively higher for tvshow and politician while the recall score is higher for government and company.

2.2 agencies

For an agency managing several verified and business Facebook pages to have as many "likes" as possible, the easiest way is to link these pages to each other. So when someone clicks on one page, he can find the links to other pages of this agency as well. Evidences are found by using the k_clique_communities method from Networkx, which looks for k-clique communities in graph using the percolation method. I set k equals 20 for tvshow and government and 10 for company and politican for a brief example. The found communities are connected subgraphs with maximal link density(complete subgraphs). Take the first such clique in tvshow as an example, we get 24: Legacy, Cosmos, Hotel Hell, Kicking Screaming, Party Over Here, etc. We will find that most of them are produced by FOX. It's highly likely that they are managed by the same agency. Further more, we may relax the requirement of clique to find more evidence for coordinated PR strategies. For agencies who want to coordinate with each other, one may link his hub page to other agencies' hub pages as well so as to gain popularity. Using greedy_modularity_communities as another way to detect communities, we get 123 communities in total with more related restriction. As an example, the first community in tvshow has 542 communities, which belong to different TV companies. Therefore, evidence shows that different agencies can work together to build mutual marketing relations and create and maintain larger following.