Uncovering Key Network Characteristics of Social Network Pages: An ERGM Analysis

Anamika Dey & Sankhadip Roy
Department of Computer Science and Statistics
University of Rhode Island, Kingston, USA

Abstract:

This article examines the network structure of a social media platform Facebook with four different types of pages: Politicians, TV shows, Government organizations, and Companies. Exponential Random Graph Models (ERGMs) are used to analyze the network structure and identify significant network statistics that explain the patterns of connections between the pages. The network is constructed from available data on the nodes and edges in the network, and various structural properties of the network are considered, including degree distributions and edge density. Goodness-of-fit tests are performed to evaluate the adequacy of the ERGM models in capturing the observed network structure. The analysis aims to gain insights into the structure of social media networks with multiple types of pages and identify the key network characteristics that shape the patterns of connections between them. The use of ERGMs provides a comprehensive understanding of the network structure and potential opportunities for research and analysis.

Introduction:

Statistical models for social networks have enabled researchers to study complex social phenomena that give rise to observed patterns of relationships among social actors and to gain a rich understanding of the interdependent nature of social ties and actors**.** la Haye et al. [5] have used Exponential-Family Random Graph Models (ERGMS) on social networks within medium to large social groups. Sadi-Nezhada et al. [4] used ERGMS to design a statistical process control through network behavior.

Social networks have an important role in today's lifestyle and detecting any changes in their structure could be vital for social systems. Change detection in a social process could be complicated because of stochastic and complex manners of its component's which are humans. For this purpose, Saghaei et al. [3] proposed a method that combines the nodal attributes with structural tendencies of ERGMs to find the best fitting model which can properly define humans' characteristics in the observed network. More literature on network analysis can be found in [1],[2] and the references therein.

Facebook is a popular social networking site that has gained widespread use globally. The platform's vast user base generates a considerable amount of data, which can be analyzed to understand various social phenomena, such as the formation of friendships, the spread of information, and the emergence of communities. In this study, we analyze a Facebook dataset to explore the patterns and characteristics of social connections among users.

We begin by conducting a basic analysis of the dataset to describe the distribution of its variables. We examine the number of users using vertices counts, the number of friendships, the density of the network, and the degree distribution. Additionally, we investigate the clustering coefficient, transitivity and assortativity to understand the network's structure.

Next, we use the Exponential Random Graph Models (ERGMs) to examine the processes that lead to the formation of social ties on Facebook. We use ERGMs to estimate the probabilities of observing certain network patterns given the underlying processes that drive the formation of social connections. We then compare these predicted network patterns to the observed network's features to assess the goodness of fit of our ERGM model.

Finally, we conduct a simulation analysis to generate alternative networks that reflect the underlying processes of social connection formation. We simulate networks using the estimated parameters of our ERGM model and compare these alternative networks to the observed network. This analysis allows us to examine how well our ERGM model captures the underlying processes of social connection formation and to generate alternative scenarios for the network's evolution.

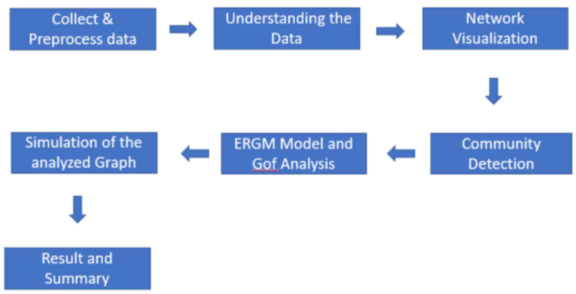
Overall, our analysis provides insights into the patterns and processes of social connection formation on Facebook, which can be used to inform future studies on online social networks.

Data Source:

https://archive.ics.uci.edu/ml/datasets/Facebook+Large+Page-Page+Network

Data Description: This web graph is a page-page graph of verified Facebook sites. Nodes represent official Facebook pages while the links are mutual likes between sites. Node features are extracted from the site descriptions that the page owners created to summarize the purpose of the site. This graph was collected through the Facebook Graph API in November 2017 and restricted to pages from 4 categories which are defined by Facebook. These categories are politicians, governmental organizations, television shows and companies.

Methodology: The study will use a network analysis approach to analyze the Facebook network data. The analysis will focus on identifying the structural properties of the network, such as centrality and community structure, and examining the relationship between these properties and visualization for different layers and analysis. The study will also detect the top 5 community and will apply the ERGM model with edge and different node factor. This model will allow us to capture the features of the network of the top communities whether the features are significant or not. Then the Goodness of Fit and simulation is applied to assess the accuracy of a statistical model that describes the network's underlying process. By doing the simulation analysis we can observe the alternative scenario of the network.



Full Network:

This Facebook network has 22,470 nodes and 171,002 undirected edges. The Edges do not have any features. The structure of the nodes is as follows:

Variable	Description
Facebook_id	Facebooks ids like 1.46e+14,1.91e+11
Page name	e.g., "The voice of China","US consulate General Mumbai" etc..
Page type	"tvshow","government", "company" and "politician"

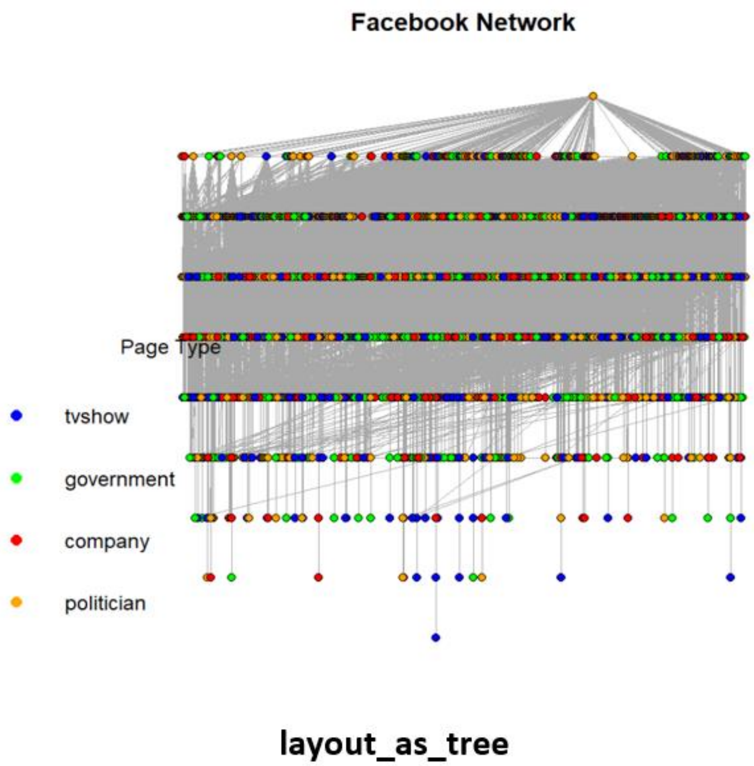


Fig 1

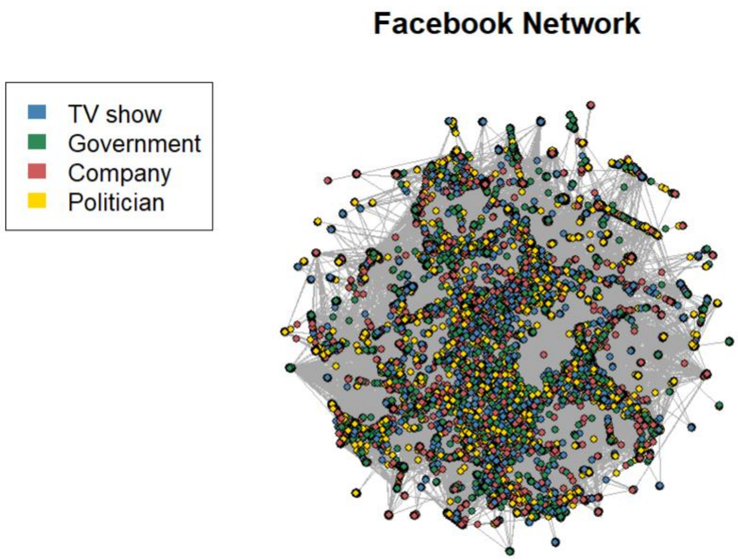


Fig 2: Facebook Network in layout.drl

Descriptive Analysis of the network:
We start with the distribution of four-page types namely, company, government, politician and tvshow and the degree distribution.

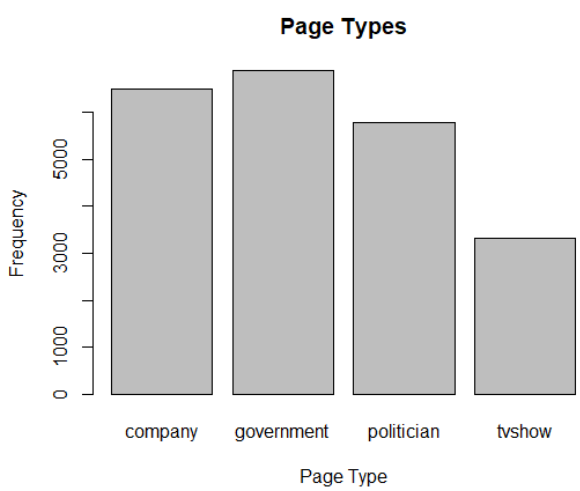


Fig 3

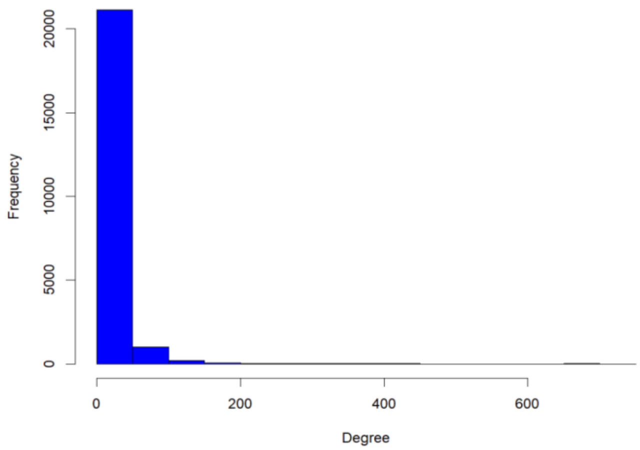


Fig 4

The histogram (Fig 4) shows that the degree distribution is highly skewed. The summary of degree distribution is below.

min	Q1	Median	mean	Q3	max	St. deviation
1	3	7	15.2	17	709	26.41

Density:

We have computed the global density and found the density of the neighbors of the first 1000 vertices which we call local densities.

Global density	Mean of local density	St dev of local density
0.00067	**0.5965**	**0.2699**

Local Density Distribution of Neighbors of the nodes

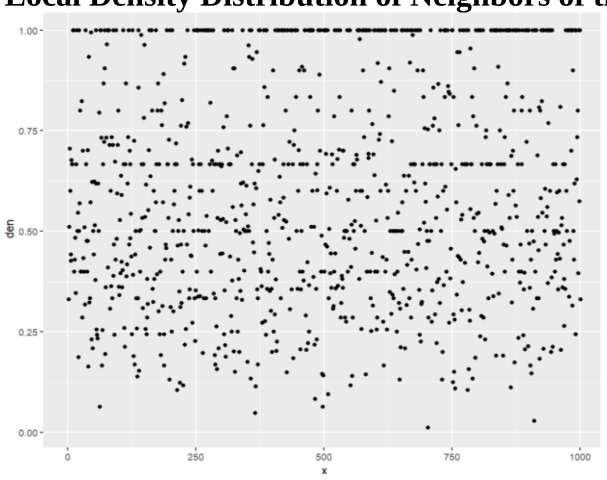


Fig 5

Although the global density of the network is very low (0.0006), but we can see in Fig 5 that the local densities among the neighbors of the vertices are reasonably high.

Clustering Coefficients: Just like density, we have also computed the global and local clustering coefficients.

Global Clustering Coef	Mean of local Clustering Coef	St dev of Clustering Coef
0.2323	**0.3485**	**0.3177**

Local Transitivity Distribution of Neighbors of the nodes

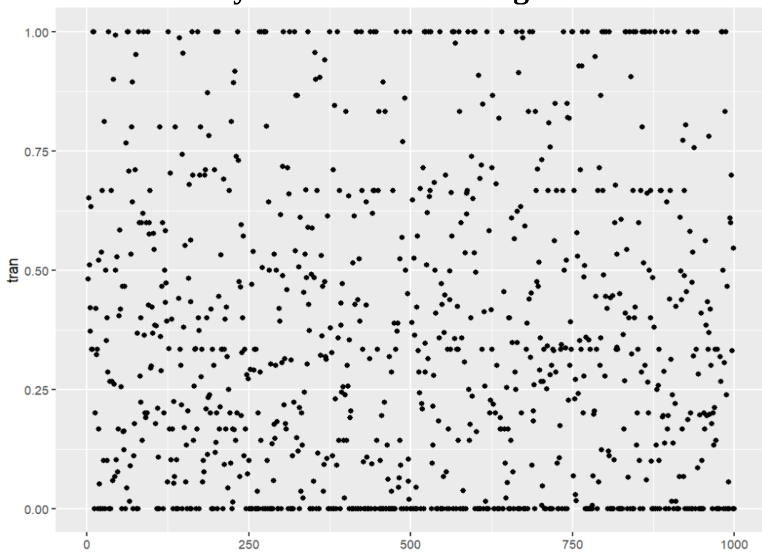


Fig 6

Many of the local clustering coefficients are more than 0.5 and some are close to 1. So, there are possibilities of multiple communities whose existence can be identified with further analysis.

Community Detection:

The basic goal of this project was to analyze the community structure of a Facebook network depending on different page types: Politicians, TV shows, Government and Companies by finding the first 5 community in the network. Our results showed that the communities were primarily composed of pages belonging to a single category, with some overlap between the government and TV show pages. We also found that certain communities had higher levels of page engagement. The reason behind it may depend on likes and shares.

ERGM Model:

Exponential-family random graph models (ERGMs) are a general class of models based on exponential-family theory for specifying the tie probability distribution for a set of random graphs or networks. Within this framework, one can—among other tasks:

• Define a model for a network that includes covariates representing features like nodal attribute homophily, mutuality, triad effects, and a wide range of other structural features of interest;

• Obtain maximum likelihood estimates for the parameters of the specified model for a given data set;

• Test the statistical significance of individual coefficients, assess models for convergence and goodness-of-fit and perform various types of model comparison; and

• Simulate new networks from the underlying probability distribution implied by the fitted model.

Here we have tried to fit ERGM on the first five communities.

Community 1 plot:

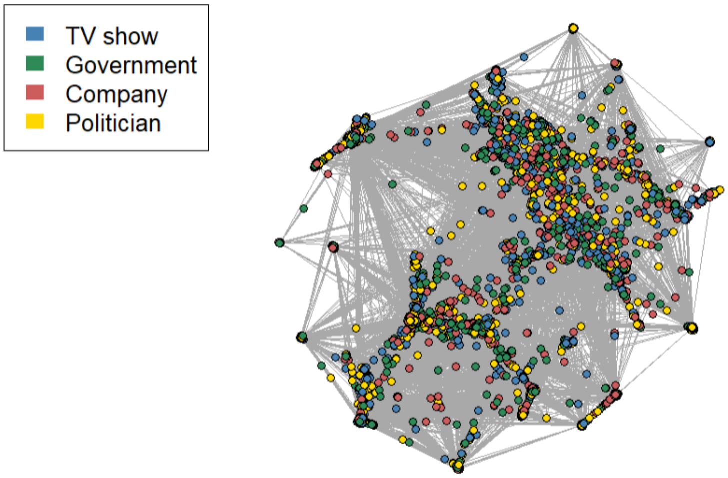


Fig 7

Distribution of page types in Community 1:

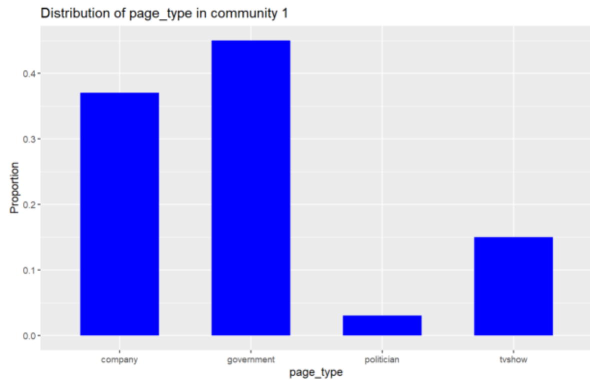


Fig 8

ERGM Model of Community 1:

Model 1:

ergm(formula = gb_1.s ~ edges + nodefactor("page_type"))

Maximum Likelihood Results:

	Estimate	Std. Error	MCMC %	z value	Pr(>|z|)
edges	-5.605902	0.010022	0	-559.372	<1e-04 ***
nodefactor.page_type.government	0.032736	0.007570	0	4.324	<1e-04 ***
nodefactor.page_type.politician	0.013604	0.007844	0	1.734	0.0829 .
nodefactor.page_type.tvshow	-0.063790	0.009438	0	-6.759	<1e-04 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null Deviance: 22177932 on 1.6e+07 degrees of freedom
Residual Deviance: 779504 on 1.6e+07 degrees of freedom

AIC: 779512 BIC: 779570 (Smaller is better. MC Std. Err. = 0)

Goodness of Fit of Model 1:

Goodness-of-fit for model statistics

	obs	min	mean	max	MC	p-value
edges	59064	59111	59408.26	59654		0.00
nodefactor.page_type.government	35928	35952	36253.91	36456		0.00
nodefactor.page_type.politician	31202	31063	31244.01	31445		0.76
nodefactor.page_type.tvshow	16805	16846	16935.29	17024		0.00

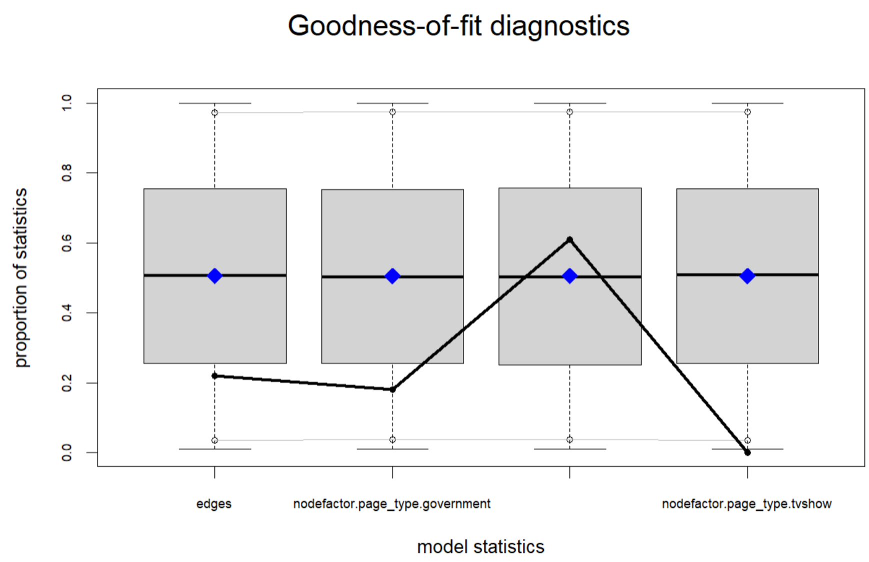


Fig 9: Goodness of Fit for model 1 of Community 1

Community 2 Plot:

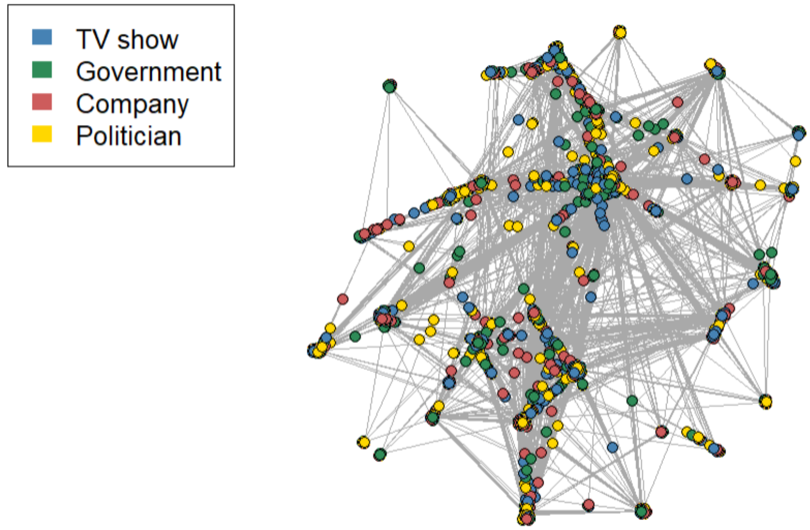


Fig 10

Distribution of page types in Community 2:

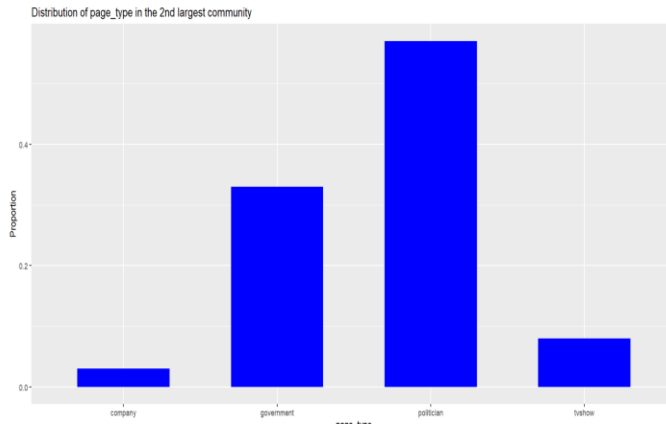


Fig 11

ERGM Model of Community 2:

Model :

ergm(formula = gb_1.s ~ edges + nodefactor("page_type"))

Maximum Likelihood Results:

	Estimate	Std. Error	MCMC %	z value	Pr(>|z|)	
edges	-5.605902	0.010022	0	-559.372	<1e-04	***
nodefactor.page_type.government	0.032736	0.007570	0	4.324	<1e-04	***
nodefactor.page_type.politician	0.013604	0.007844	0	1.734	0.0829	.
nodefactor.page_type.tvshow	-0.063790	0.009438	0	-6.759	<1e-04	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null Deviance: 22177932 on 1.6e+07 degrees of freedom
Residual Deviance: 779504 on 1.6e+07 degrees of freedom

AIC: 779512 BIC: 779570 (Smaller is better. MC Std. Err. = 0)

Goodness of Fit of the model:

Goodness-of-fit for model statistics

	obs	min	mean	max	MC p-value
edges	59064	58803	58962.92	59154	0.20
nodefactor.page_type.government	35928	35576	35704.71	35883	0.00
nodefactor.page_type.politician	31202	31081	31241.03	31396	0.88
nodefactor.page_type.tvshow	16805	16707	16855.20	16970	0.42

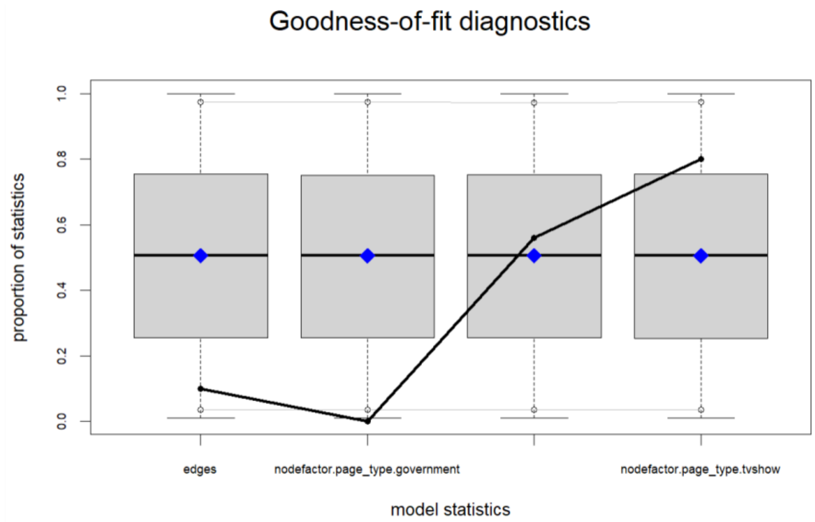


Fig 12: Goodness of Fit for model of Community 2

Community 3 plot:

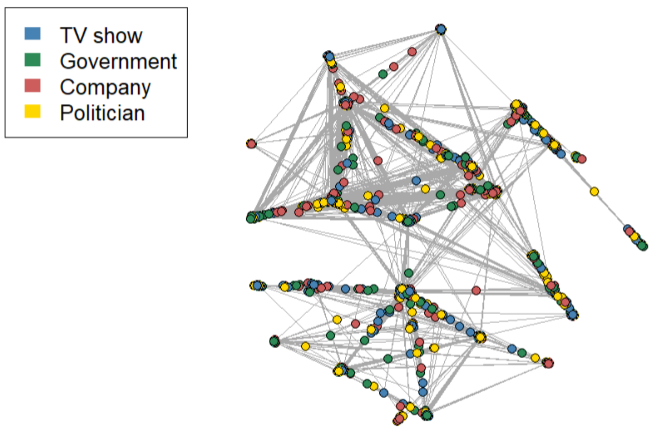


Fig 13

Distribution of page types in Community 3:

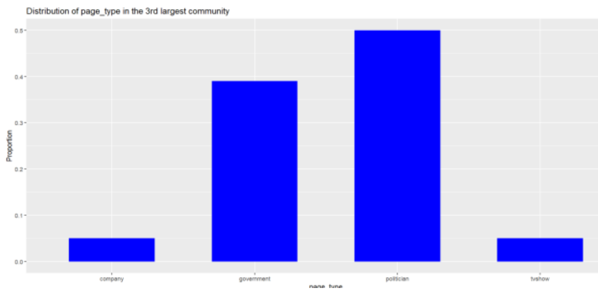


Fig 14

ERGM Model:

ergm(formula = gb_3.s ~ edges + nodefactor("page_type"))

Maximum Likelihood Results:

	Estimate	Std. Error	MCMC %	z value	Pr(>|z|)	
edges	-4.757197	0.015680	0	-303.401	< 1e-04	***
nodefactor.page_type.government	-0.009597	0.011929	0	-0.805	0.421106	
nodefactor.page_type.politician	-0.033490	0.012599	0	-2.658	0.007857	**
nodefactor.page_type.tvshow	-0.050942	0.014779	0	-3.447	0.000567	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null Deviance: 3964297 on 2859636 degrees of freedom
Residual Deviance: 272046 on 2859632 degrees of freedom

AIC: 272054 BIC: 272106 (Smaller is better. MC Std. Err. = 0)

Goodness of Fit:

Goodness-of-fit for model statistics

	obs	min	mean	max	MC	p-value
edges		23457	23153	23322.65	23518	0.22

nodefactor.page_type.government	14391	14151	14264.84	14382	0.00
nodefactor.page_type.politician	11657	11448	11558.58	11714	0.24
nodefactor.page_type.tvshow	6897	6855	6935.23	7035	0.58

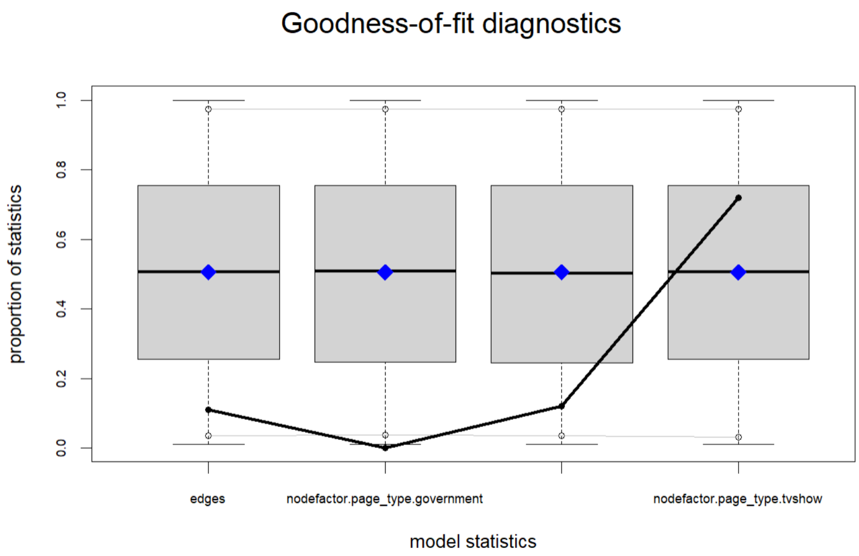


Fig 15

Community 4 plot:

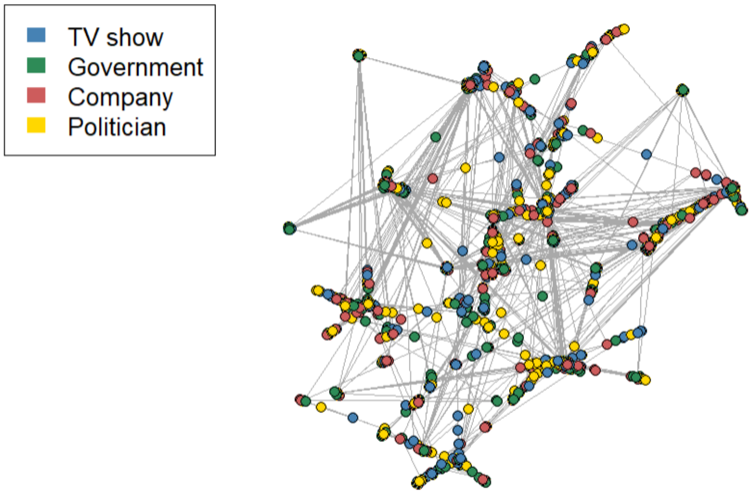


Fig 16

Distribution of page types in Community 4:

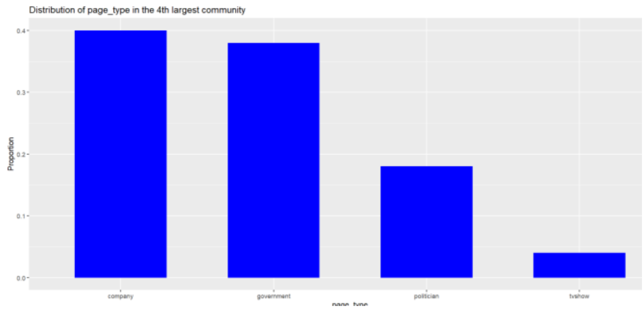


Fig 17

Community 5 Plot:

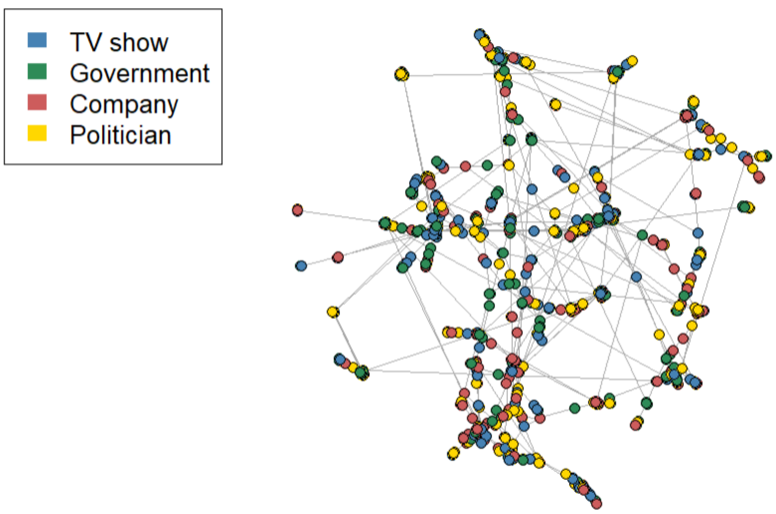


Fig 18

Distribution of page types in Community 5:

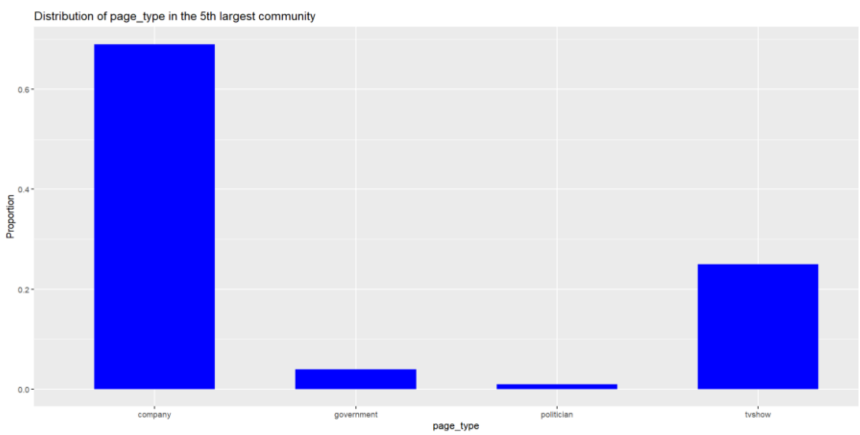


Fig 19

Network Type	Vertices	Assortativity	Transitivity
Full Network	22,470	0.0849125	0.2323214
Community 1	5657	0.02802067	0.2251863
Community 2	4009	0.02293722	0.2568745
Community 3	2392	0.05837795	0.3415304
Community 4	2314	0.07972632	0.3129261
Community 5	1279	0.1483445	0.3863182

Simulating networks:

Another notable feature of ERGMs is that they are generative. We can simulate networks that are near the maximum likelihood realization of sufficient statistics. This can be useful for examining fit, among other things, and is easy using the S3 method for simulate for an ERGM object. In addition to checking model fit, we can change parameter values, constrain the network in various ways.

To add a term for homophily within the community we use the term **nodematch**, which takes at least one argument (the nodal attribute) and provides the change in the likelihood of a tie if the two nodes match on that attribute. Note that we can estimate a differential homophily effect the change in tie likelihood for two nodes being in the same group can vary by group.

Simulated Graph of Community 1:

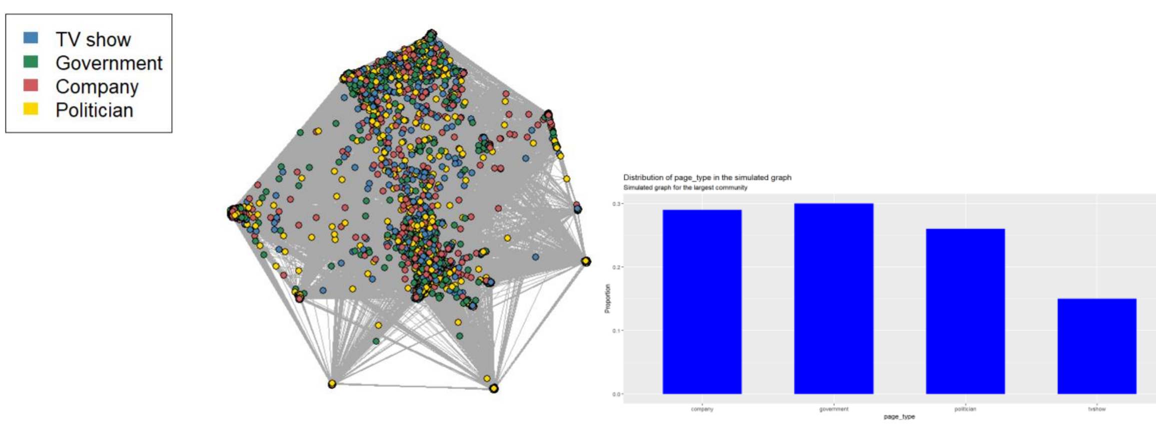


Fig 20

Simulated Graph of Community 2:

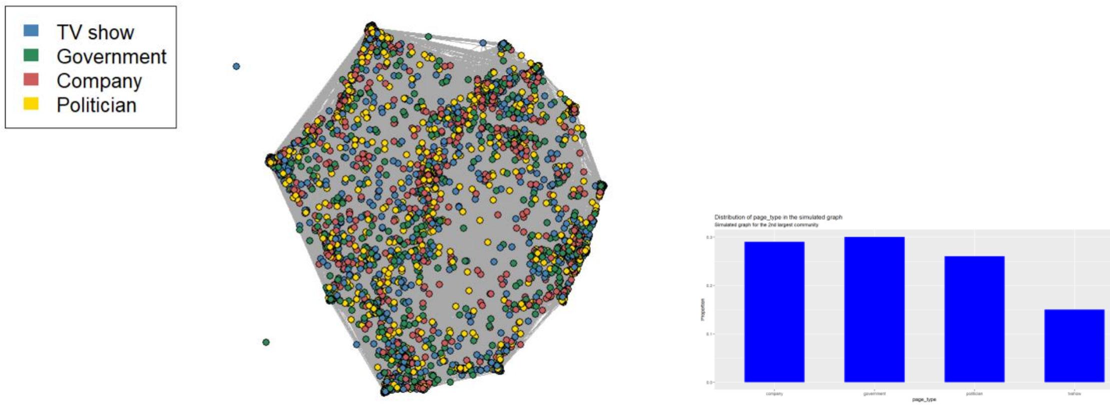


Fig 21

Simulated Graph of Community 3:

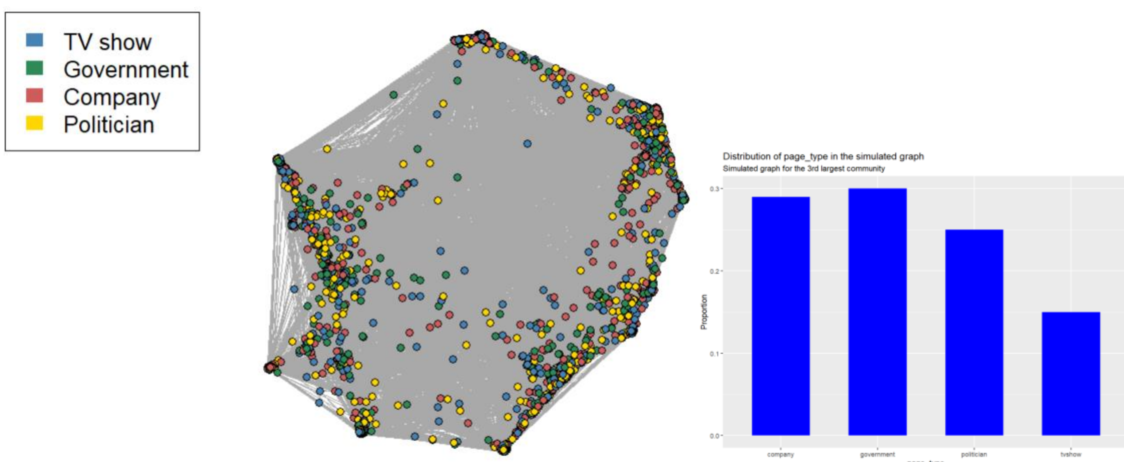


Fig 22

Simulated Graph of Community 4:

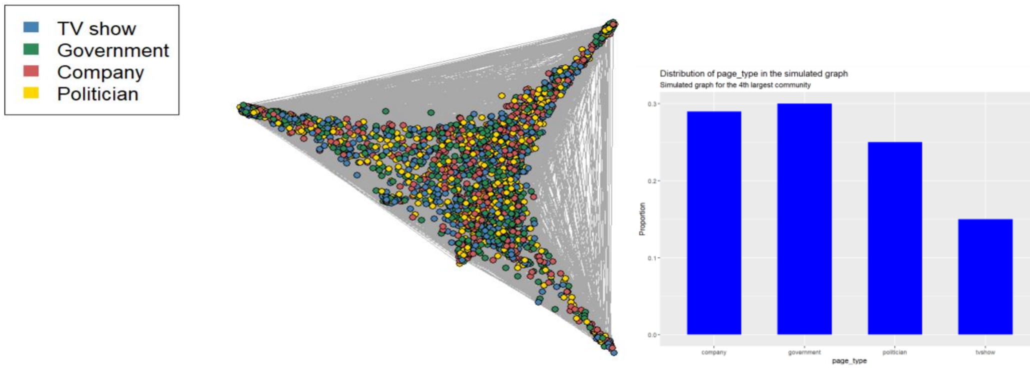


Fig 23

Network Type	Vertices	Assortativity	Transitivity
Full Network	22,470	0.0849125	0.2323214
Simulation of Community 1	5657	0.1051087	0.120598
Simulation of Community 2	4009	0.05798483	0.1494928
Simulation of Community 3	2392	0.06901998	0.142986
Simulation of Community 4	2314	0.1143001	0.01120581

Conclusion and possibilities of future work:

In this article, we have analyzed a Facebook Social Network. We have 22,470 facebook ids and 171,002 connections between them. The ids follow certain popular topics like Television Shows, Companies, Politicians and Governments. The existence of different communities was quite evident from the descriptive analysis. We have analyzed the first five communities and we have found different distributions of page types in each community. We have fitted ERGM for those communities against the page types and the summary analysis and goodness of fit have shown the significance of different page types in those models. Finally, we have generated simulated graphs from those models for each community. This network can further be analyzed. New analysis like predicting links between non-adjacent nodes depending on the common page types can be quite significant.

References:

[1] B. Rozemberczki, C. Allen and R. Sarkar. Multi-scale Attributed Node Embedding. 2019.

[2] Wang, P., Robins, G., Matous, P. (2016). Multilevel Network Analysis Using ERGM and Its Extension. In: Lazega, E., Snijders, T. (eds) Multilevel Network Analysis for the Social Sciences. Methodos Series, vol 12. Springer, Cham.

[3] S. Golshid Sharifnia & Abbas Saghaei (2022) A statistical approach for social network change detection: an ERGM based framework, Communications in Statistics - Theory and Methods, 51:7, 2259-2280.

[4] Farshid Rajabia, Abbas Saghaeia, Soheil Sadi-Nezhada, Monitoring of Social Network and Change Detection by Applying Statistical Process: ERGM, Journal of Optimization in Industrial Engineering, Vol.13, Issue 1, Winter & Spring 2020, 131-143.

[5] George G. Vega Yon, Andrew Slaughter, Kayla de la Haye, Exponential random graph models for little networks, Social Networks, Volume 64, 2021, Pages 225-238.