Group 51 – Data Mining I Project

Analysis of Social Circles and Anomalous Users in the Friendfeed Network

Jonas Blum

Sanjatul Islam

Xin Tian

Riccardo Rebecchi

Abdulfattah Morad

Research Questions

- Is it possible to identify distinct social circles within the Friendfeed network based on user connections, and characterize them in terms of shared interests and activity patterns?

- Is it possible to detect anomalous users in the Friendfeed network (e.g., bots, spammers) based on posting, following, and interaction patterns (commenting/liking)?

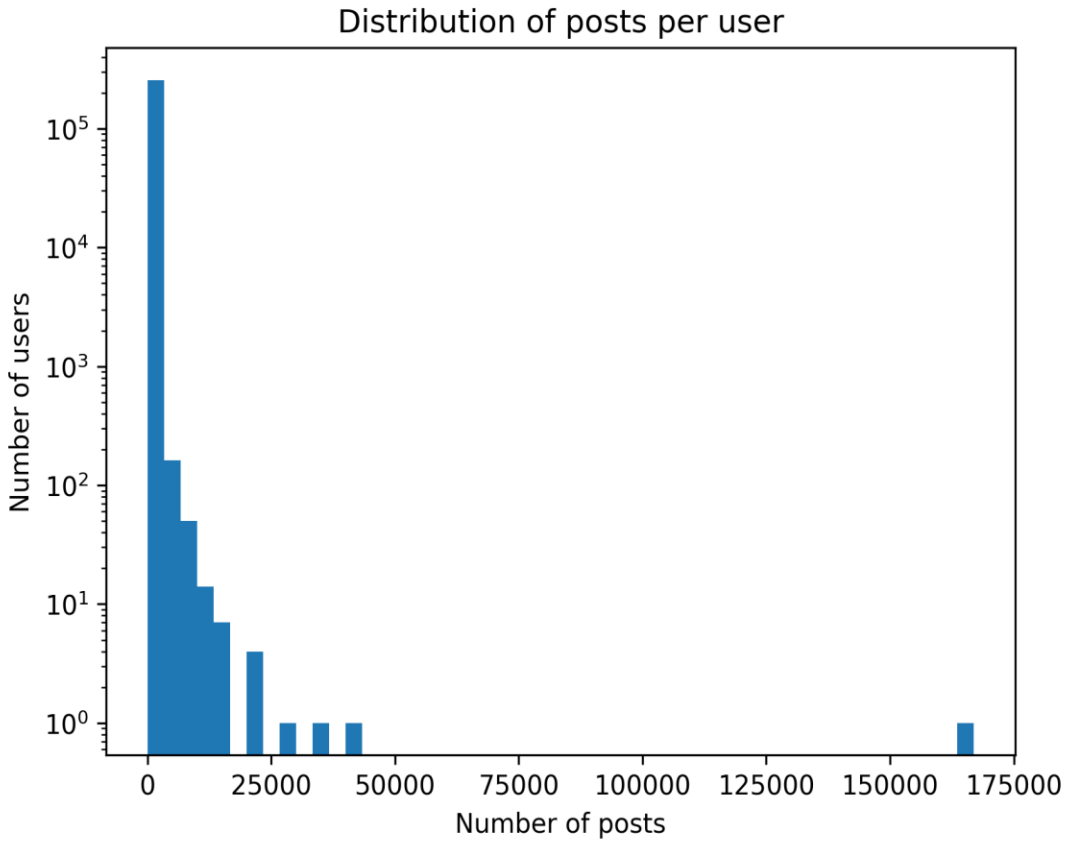
Data – Friendfeed Social Network

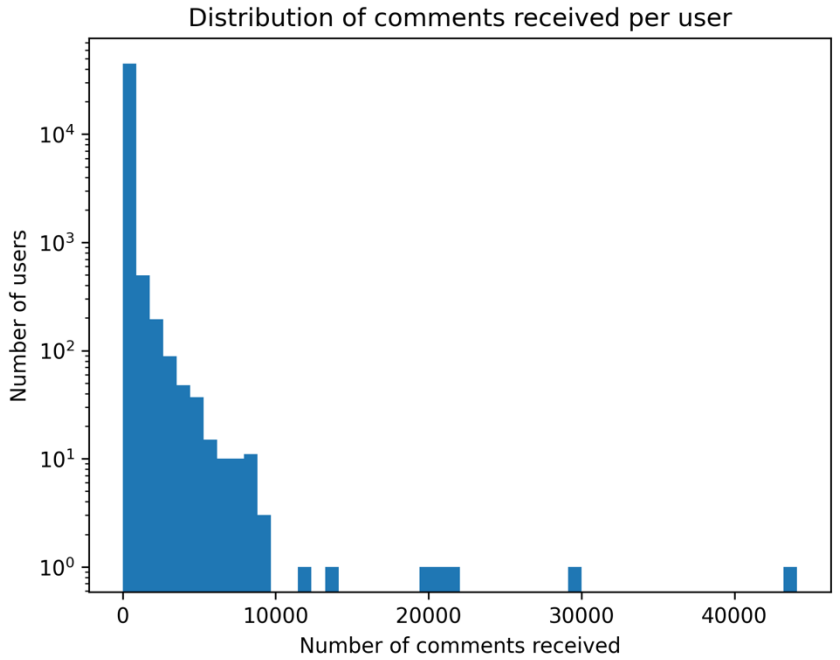
- Number of users: 665’382
- Number of posts: 12’450’658
- Number of comments: 3’749’891
- Number of people following: 19’547’158
- Number of likes: 798’112

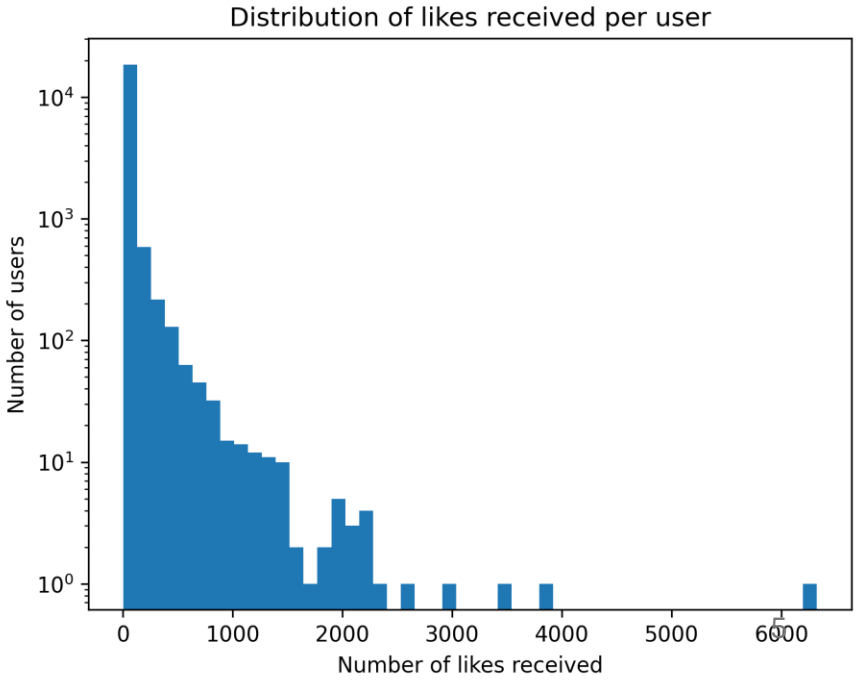
Preprocessing

- Dataset has already been preprocessed before
- Remove “dead” accounts (without any posts, likes and comments)
  Users removed: 160’792 (504’590 users remaining)
- Connect users following each other (via FollowedID and FollowerID)
- Connect users and their posts via UserID
- Connect posts, users and likes + comments via PostId and UserID

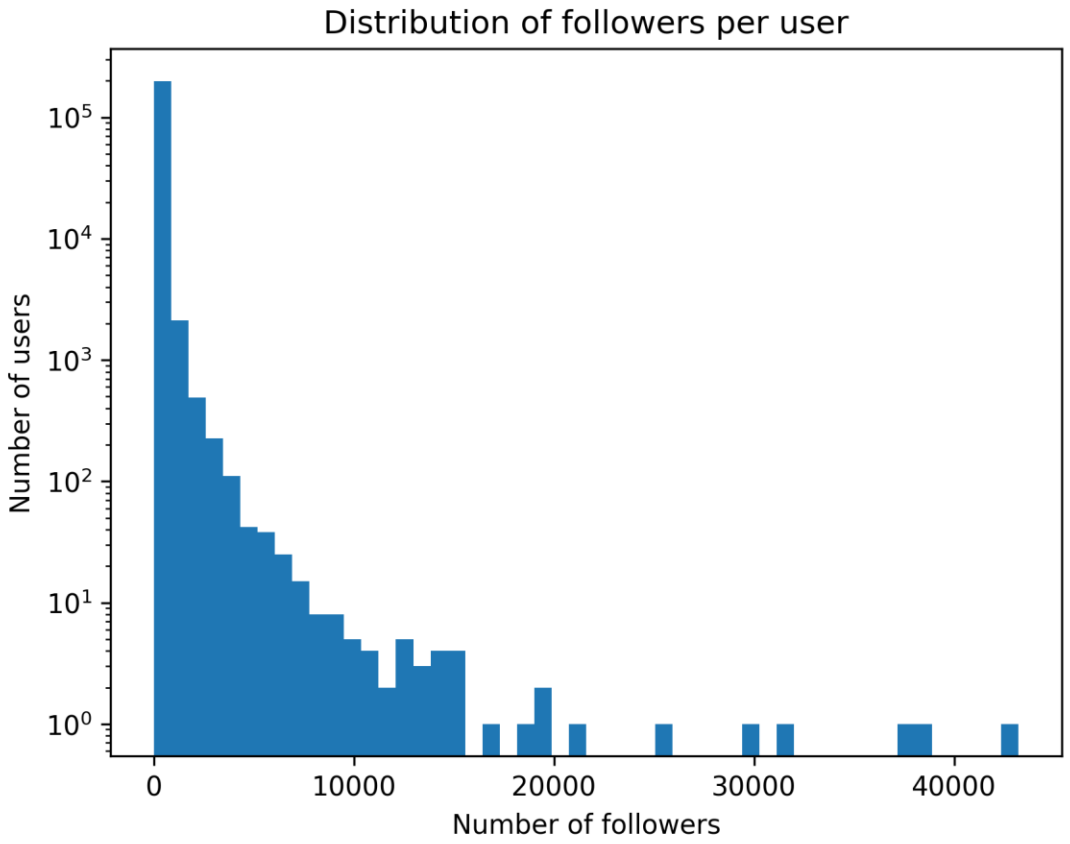
Initial Data Exploration

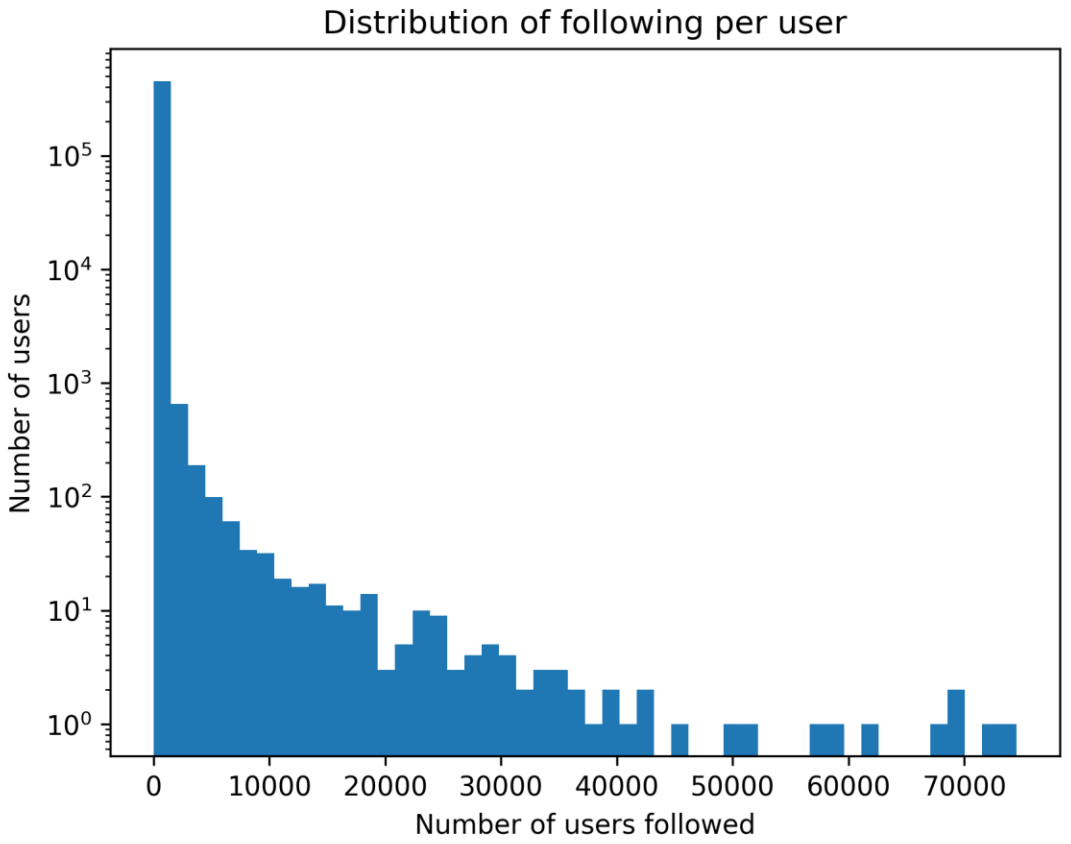






Initial Data Exploration





Initial Data Exploration – per User Statistics

Type	Average	Median	Standard Deviation	Maximum
Follower Count per User	96.7	20	380.44	43’222
Following Count per User	43.04	4	534.28	74’522
Posts Created per User	48.53	8	433.89	166’965
Likes Received per Post per User	0.04	0	0.63	180.7
Likes Given per User	34.89	3	213.05	25’224
Comments Received per Post per User	0.2	0	1.69	279.15
Comments Given per User	69.38	4	339.45	29’348

Louvain Algorithm for Social Circle Detection

- 504’590 users/vertices
- 19’547’158 followings/edges
- Invented in 2015
- Space and time complexity: O(|V| + |E|)
- Similar to agglomerative clustering: bottom-up, hierarchical and greedy

Community 6 (157’643 Users)

Type	Average	Median	Standard Deviation	Maximum
Follower Count per User	109.66	36	491.91	42487
Following Count per User	38.76	4	570.49	58349
Posts Created per User	31.86	8	191.34	22375
Likes Received per Post per User	0.01	0	0.13	16
Likes Given per User	33.95	2	165.45	3948
Comments Received per Post per User	0.07	0	0.34	34
Comments Given per User	24.76	3	138.02	4365

Community 1 (84’830 Users)

Type	Average	Median	Standard Deviation	Maximum
Follower Count per User	44.16	12	364.06	36765
Following Count per User	25.48	3	681.91	73103
Posts Created per User	59.5	9	326.26	16258
Likes Received per Post per User	0	0	0.06	5
Likes Given per User	9.65	1	43.66	666
Comments Received per Post per User	0.05	0	0.21	14
Comments Given per User	48.15	3	307.91	12197

Community 0 (48’550 Users)

Type	Average	Median	Standard Deviation	Maximum
Follower Count per User	304.79	135	566.65	10232
Following Count per User	143.47	13	645.94	26764
Posts Created per User	65.25	10	410.25	40377
Likes Received per Post per User	0	0	0.06	3
Likes Given per User	7.23	1	34.45	419
Comments Received per Post per User	0.04	0	0.18	7
Comments Given per User	27.52	3	162.12	4644

Community 9 (46’322 Users)

Type	Average	Median	Standard Deviation	Maximum
Follower Count per User	88.18	19	237.63	7997
Following Count per User	26.37	3	150.25	18364
Posts Created per User	61.65	7	1472.32	166965
Likes Received per Post per User	0.23	0	1.08	78
Likes Given per User	34.89	4	328.76	25224
Comments Received per Post per User	0.87	0	2.24	49
Comments Given per User	141.22	8	574.59	29348

Community 5 (25’606 Users)

Type	Average	Median	Standard Deviation	Maximum
Follower Count per User	40.75	13	121.29	7044
Following Count per User	29.61	8	89.54	7098
Posts Created per User	58.23	23	140.39	6949
Likes Received per Post per User	0	0	0.02	1
Likes Given per User	8.96	2	18.98	133
Comments Received per Post per User	0.01	0	0.1	2
Comments Given per User	23.29	4	126.62	3773

Algorithms used for Anomaly Detection

- Statistical Approach (Multivariate Gaussian)
- Clustering-based Approach (k-Means)
- Distance-based Approach (k-NN)
- Density-based Approach (LOF)
- Isolation-based Approach (iForest)
- One-Class SVM
- Reconstruction-based Approach (PCA)

Features used for Anomaly Detection

- Follower Count per User
- Following Count per User
- Posts Created per User
- Likes Received per Post per User
- Likes Given per User
- Comments Received per Post per User
- Comments Given per User

Outlier User ”pattonroberta”

Type	Value	Rank	Top in %
Follower Count	1,453	1280	0.26
Following Count	147	19283	3.97
Posts Created	1,389	793	0.16
Likes Received per Post (on average)	0.66	2992	0.61
Likes Given	2,707.0	5	0.00
Comments Received per Post (on average)	3.28	2013	0.41
Comments Given	4,365.0	36	0.01

Algorithms that detected outlier: Multivariate Gaussian, kNN, iForest, PCA Reconstruction

Outlier User ”golaqa”

Type	Value	Rank	Top in %
Follower Count	0	196437	70.13
Following Count	63	42306	8.72
Posts Created	215	8210	1.69
Likes Received per Post (on average)	1.67	467	0.1
Likes Given	3,439.0	4	0
Comments Received per Post (on average)	14.07	129	0.03
Comments Given	4,259.0	38	0.01

Algorithms that detected outlier: Multivariate Gaussian, kMeans, kNN, PCA Reconstruction

Outlier User ”jluvisions”

Type	Value	Rank	Top in %
Follower Count	110	36065	7.42
Following Count	0	443553	95.45
Posts Created	8,297.0	38	0.01
Likes Received per Post (on average)	0	15716	3.22
Likes Given	0	15739	51.61
Comments Received per Post (on average)	0.99	8591	1.76
Comments Given	8,205.0	5	0

Algorithms that detected outlier: Multivariate Gaussian, kMeans, kNN, PCA Reconstruction

Outlier User ”romanussum”

Type	Value	Rank	Top in %
Follower Count	53	60294	12.42
Following Count	2	263137	59
Posts Created	12,198.0	16	0
Likes Received per Post (on average)	0	15915	51.63
Likes Given	0	15739	51.61
Comments Received per Post (on average)	1	8494	1.74
Comments Given	12,197.0	2	0

Algorithms that detected outlier: Multivariate Gaussian, kMeans, kNN, PCA Reconstruction

Outlier User ”tahaakcakaya”

Type	Value	Rank	Top in %
Follower Count	287	13630	2.8
Following Count	263	9830	2.02
Posts Created	186	9842	2.02
Likes Received per Post (on average)	0.61	3200	0.66
Likes Given	1,447.0	21	0
Comments Received per Post (on average)	25.62	28	0.01
Comments Given	11,546.0	3	0

Algorithms that detected outlier: Multivariate Gaussian, kMeans, kNN, PCA Reconstruction

Future Work

- Characterize the different social circles based on the data (e.g. social circle of influencers)
- Characterize the outlier users based on the data (e.g. bot, spammer, influencer)
- Highlight why the characterization of the social circle/outlier user was done