POKEC Online Social Network

1. Curation Rationale

Authors: Dhivya Eswaran, Srijan Kumar, Christos Faloutsos [1]. The dataset is built on top of [2].
Purpose: Study higher order label homogeneity [1].
Domain: Online social network
Contents: 66% of all the users and the friendship relationships on the POKEC platform
Node and Edge Semantics: Each node represents a user account. Each edge represents friendship relationship between user accounts.

2. Dataset Collection, Preprocessing and Annotation

2.1 Data Collection

Data collection mechanism:
1. Insert user into queue. The user is identified by the nick name.
2. Take the first nick from the queue. If the queue is empty, algorithm ends.
3. Take a profile by using Pokec's URL together with the nick added to

Network sampling: The data contains ~66% of all the users on the POKEC platform. Users are crawled in a breadth-first search manner, so the network is

2.2 Data Preprocessing

Network construction: Not exist
Data cleaning: Not exist
Data filtering: Not exist
Network transformation: Not exist
Attribute transformation: Anonymize the name of the users. Extract the region information from user profile by only considering the 8 regions in Slovekia, Czech, and abroad. If the user profile information is unknown, assign the label 'banskobystricky' (the second most popular region in the dataset)
Data splits: Not exist

2.3 Instance Demographics

POKEC is an online platform in Slovekia, and it is in Slovak lanugage. 89.6% of users are from Slovekia, 1.9% of users are from Czech, and 8.5% of users are from other countries.

2.4 Data Annotation

Not exist

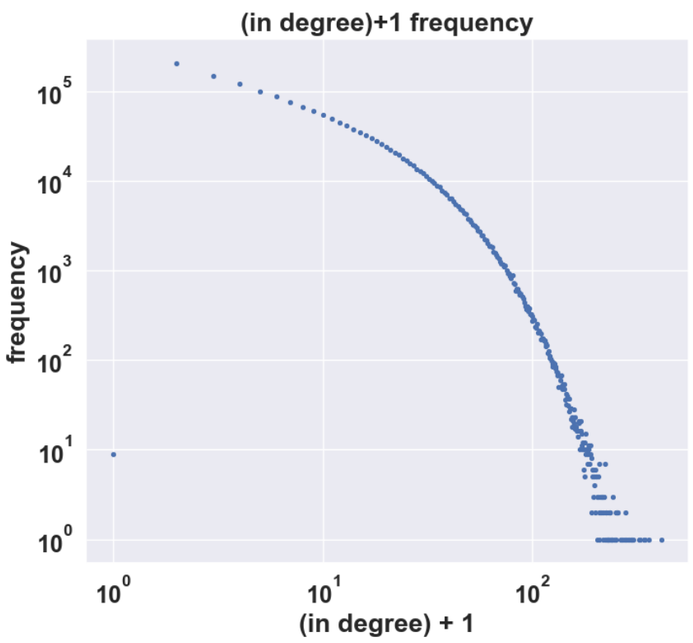
3. Uses

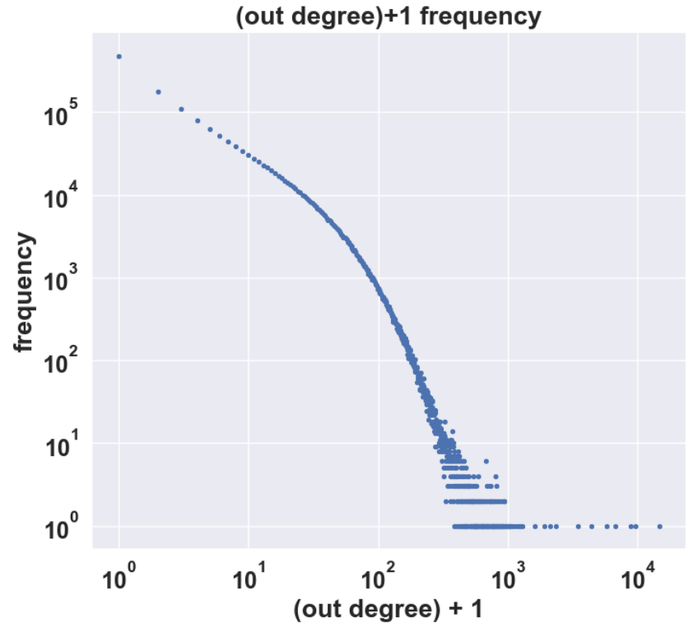
Primary intended uses: Studying higher-order label homogeneity and spreading in graphs.
Other uses: Modeling online friendship behavior.

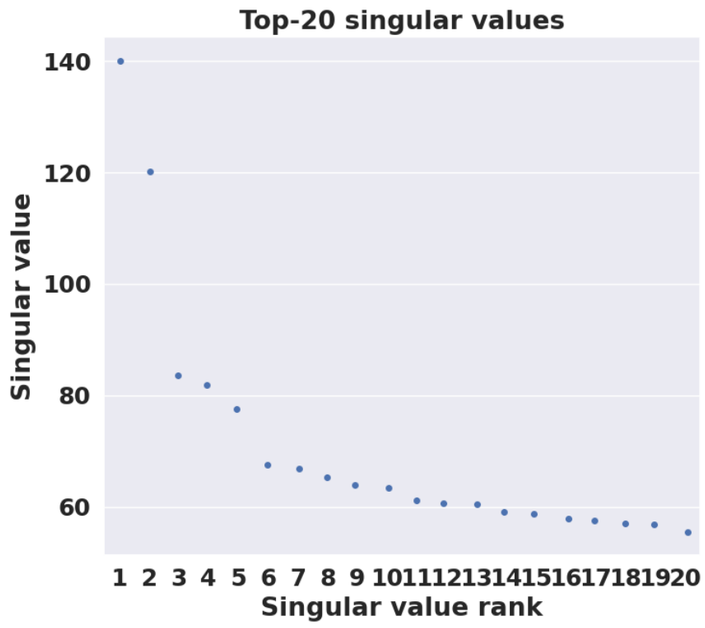
4. Network Statistics

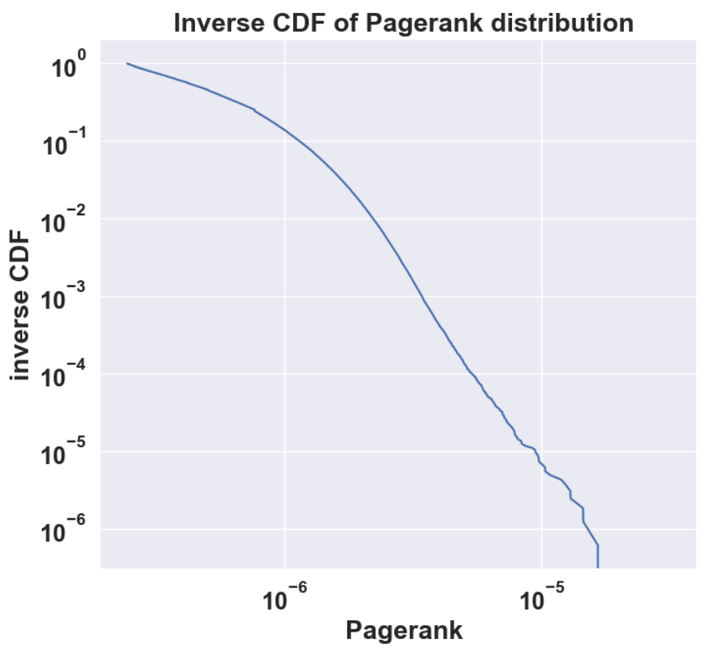
Table 1 Point Statistics

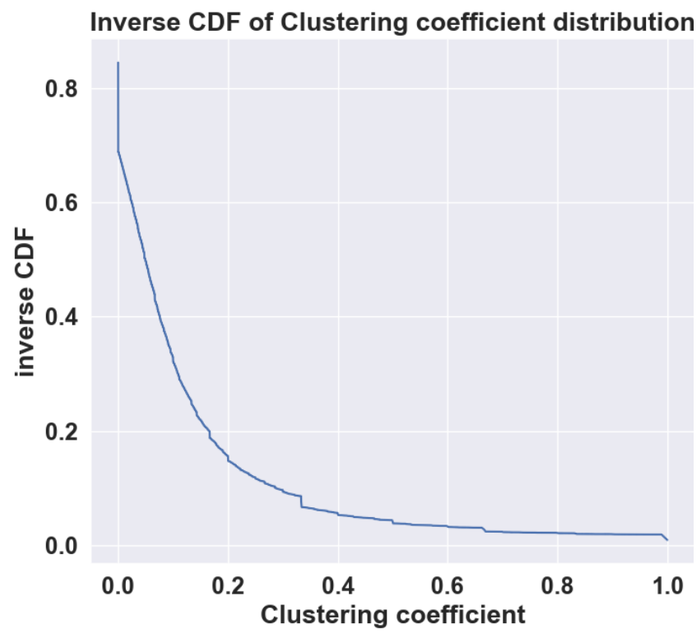
Size of nodes	1,632,803	Avg. triangle count	59.81
Size of edges	30,622,564	Avg. clustering coef.	0.11
Avg. in (out) degree	13.66	Assortativity	-0.016
% in LCC	100%	Algebraic connectivity	0.018
Max k-core	47	Spectral Radius	11.59
(Tail) power law exponent	1.40		

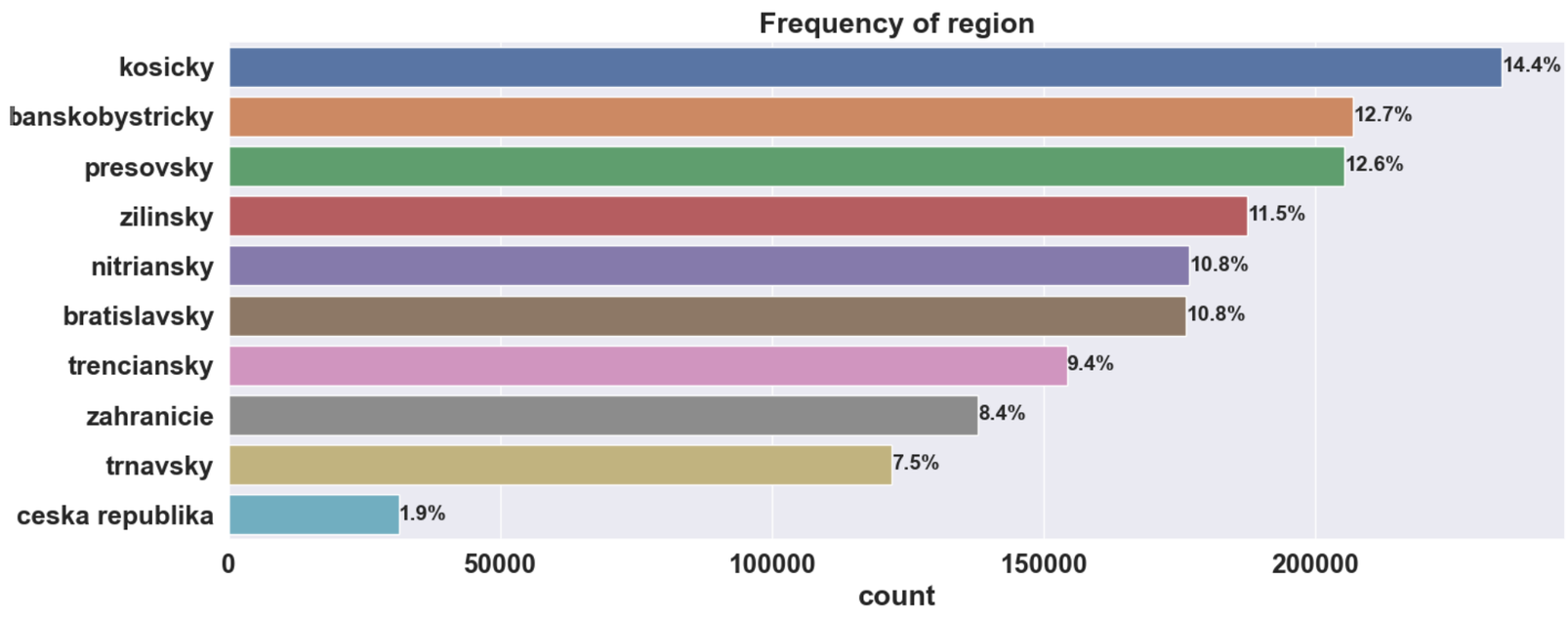












[1] Eswaran, D., Kumar, S., & Faloutsos, C. (2020). Higher-Order Label Homogeneity and Spreading in Graphs. Proceedings of The Web Conference 2020.
[2] Takac, L. (2012).Data Analysis in Social Networks.