Theory and Methods for the Analysis of Social Networks

Instructor: Alexander Volfovsky
Department of Statistical Science, Duke University

Applied paper

The structure of online social networks mirrors those in the offline world

Paper by Dunbar, Arnaboldi, Conti and Passarella (Social Networks)

1. Applied area: Psychology
2. Type of data: online social networks (turned into ego networks)
3. Goal: compare online network "layer" structure to offline structure.
4. Analytic tool: Cluster the frequency of contact of each ego network to search for layered structure.

Background

- "Social brain hypothesis" (central cognitive constraint)

Background

- "Social brain hypothesis" (central cognitive constraint)
  - Typical size of social groups correlates with the size of the neocortex.

Background

- "Social brain hypothesis" (central cognitive constraint)
  - Typical size of social groups correlates with the size of the neocortex.
  - Notion of information capacity.

Background

- "Social brain hypothesis" (central cognitive constraint)
  - Typical size of social groups correlates with the size of the neocortex.
  - Notion of information capacity.
  - Evidence from neuroimaging studies.

Background

- "Social brain hypothesis" (central cognitive constraint)
  - Typical size of social groups correlates with the size of the neocortex.
  - Notion of information capacity.
  - Evidence from neuroimaging studies.
- (Aside: mentalising?)

Background

- "Social brain hypothesis" (central cognitive constraint)
  - Typical size of social groups correlates with the size of the neocortex.
  - Notion of information capacity.
  - Evidence from neuroimaging studies.
- (Aside: mentalising?)
- "Normal"(??) social structure has layers of sizes 5, 15 and 50.

Where to find data?

- **EuroSys '09 Datasets**
  Following our EuroSys Facebook measurement study, we are making some datasets of social graphs and interaction graphs available.

  These graphs only contain simple edges connecting anonymized nodeIDs. The social graph file is simply a list of all edges in the graph, each bidirectional edge represented by a two-tuple of anonymized nodeIDs. Our user connectivity graphs reflect measurements performed in early 2008, and are not reflective of current Facebook topologies.

  For the anonymized interaction graphs, we filter interactions based on their relative age to the time of the crawl (April 2008). Each edge in the interaction graph is listed in the file as a two-tuple of anonymized nodeIDs. The interaction graph is an undirected graph, so an edge from **A** to **B** represents a bidirectional edge connecting them. We include multiple interactions within the same period as duplicate edges across the same endpoints to account for user pairs that interact more than once during the time period. This frequency can be used to assign "weights" to edges on the interaction graph. If you want an undirected, unweighted interaction graph, then remove those duplicate edges.

  Note: If you would like access to this data, please send email to *ravenben at cs dot ucsb dot edu*. When you get access to the data files, please do not distribute them beyond your immediate research group. Thank you.

  - Anonymized social graphs
    - Anonymous regional network **A**: 3,097,166 users, 28,377,481 edges, 605MB GZipped
    - Anonymous regional network **B**: 2,937,614 users, 24,236,701 edges, 521MB GZipped
  - Anonymized interaction graphs
    - Anonymous regional network **A**, n=1, t=1 month: 1,412,252 interactions, 22MB GZipped
    - Anonymous regional network **A**, n=1, t=1/2 year: 7,483,904 interactions, 87MB GZipped
    - Anonymous regional network **A**, n=1, t=1 year: 16,889,111 interactions, 159MB GZipped
    - Anonymous regional network **A**, n=1, t=lifetime: 17,644,327 interactions, 164MB GZipped

Methodology

Data set 1

- collected from Facebook pre-2009 when users within "regional" network have complete access
- Covers $\sim 56\%$ of Facebook profiles (3 million) and $\sim 37\%$ friendships (23 million).
- How were the data collected:
  - crawler obtained COMPLETE public profiles.
  - followed all friendship links
  - if privacy settings too high, profiles were not downloaded but friendships were noted.
- What's included: four time periods when contact could have been made

Methodology

Data set 1, continues

- "Active" relationship requires at least one interaction
- "Intimacy" is measured by contact frequency within a time period
- For analysis they only use people with > 10 interactions per month
- Final data are 130k egos and 5.3million edges.
- Most ego networks are smaller than 100
- Missing data: posts from public profiles to non-public profiles and between non-public profiles.
- Imputation: randomly selected 44% of nodes and assumed that those are non-public. Double the number of interaction on all the links of the ego networks of those nodes.

Facebook contact info

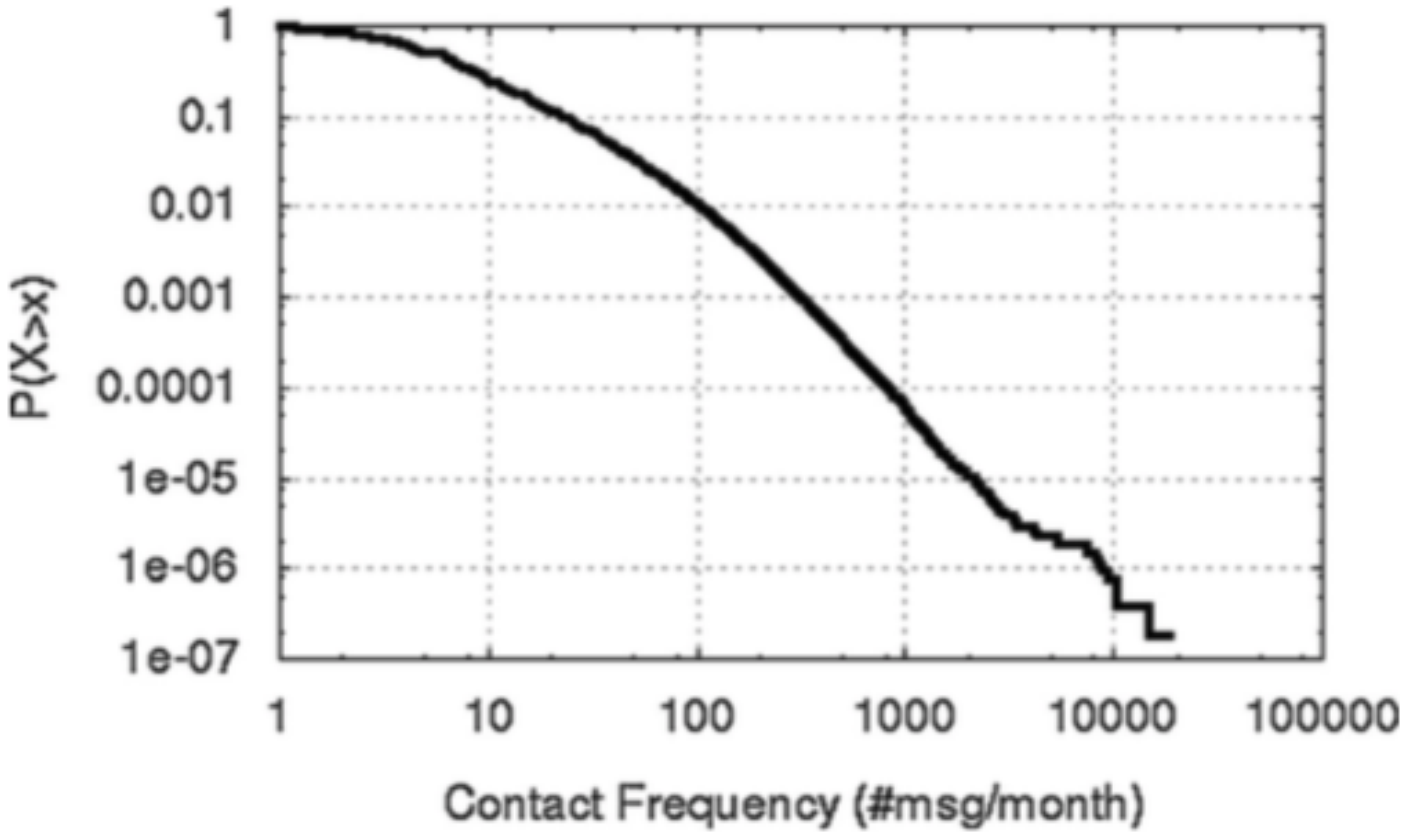

Fig. 1. CCDF of the contact frequency for relationships in Facebook dataset #1.

Facebook contact info

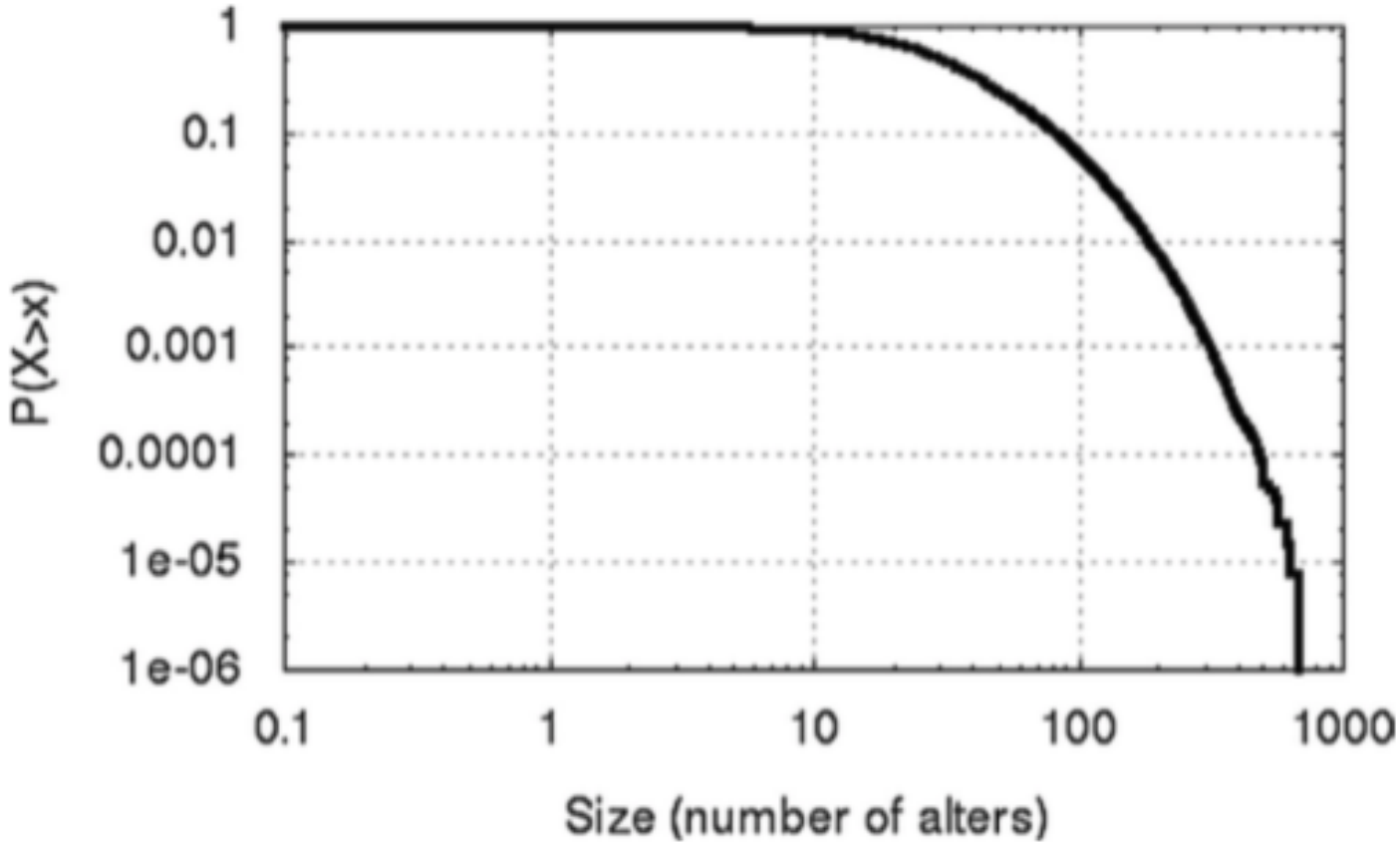

Fig. 2. CCDF of the size of ego networks for relationships in Facebook dataset #1.

Facebook contact info

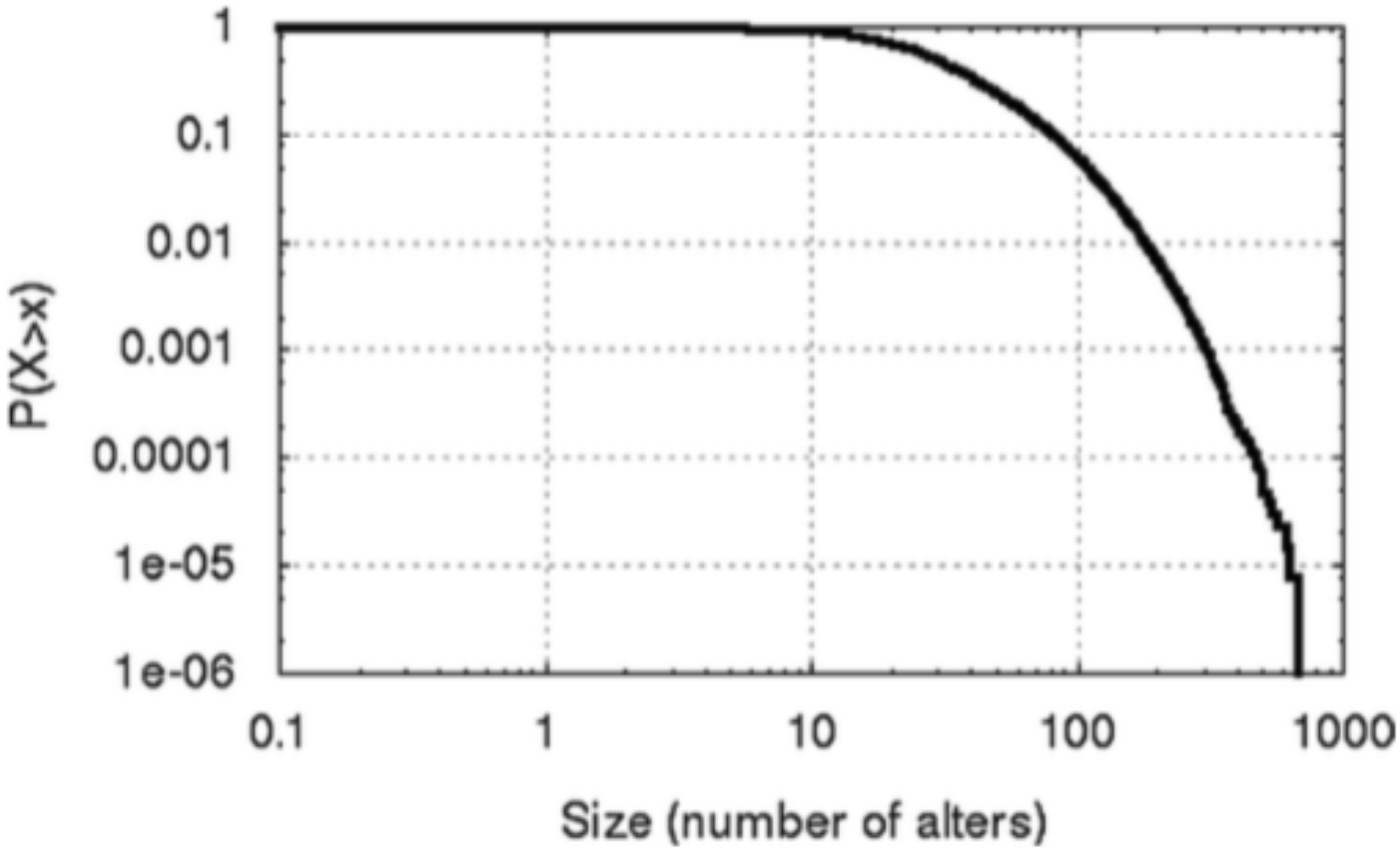

Fig. 3. CCDF of the size of ego networks considering only relationships with contact frequency higher than one message per year (active network) in Facebook dataset #1.

Methodology

Data set 3

- collected from Twitter (303k user profiles) in November 2012.
- Looks at "mentions" and "replies" (direct communication)
- Use only these types of data to measure "intentionality" in the communications.
- Frequency of contact is measured by

$$f(u_1, u_2) = \frac{N_{rep}(u_1, u_2)}{d(u_1, u_2)}$$

  where N is the number of replies from u_1 to u_2 and d is the duration of the relationship between them.
- Data are filtered for "human behavior"
-

Facebook contact info

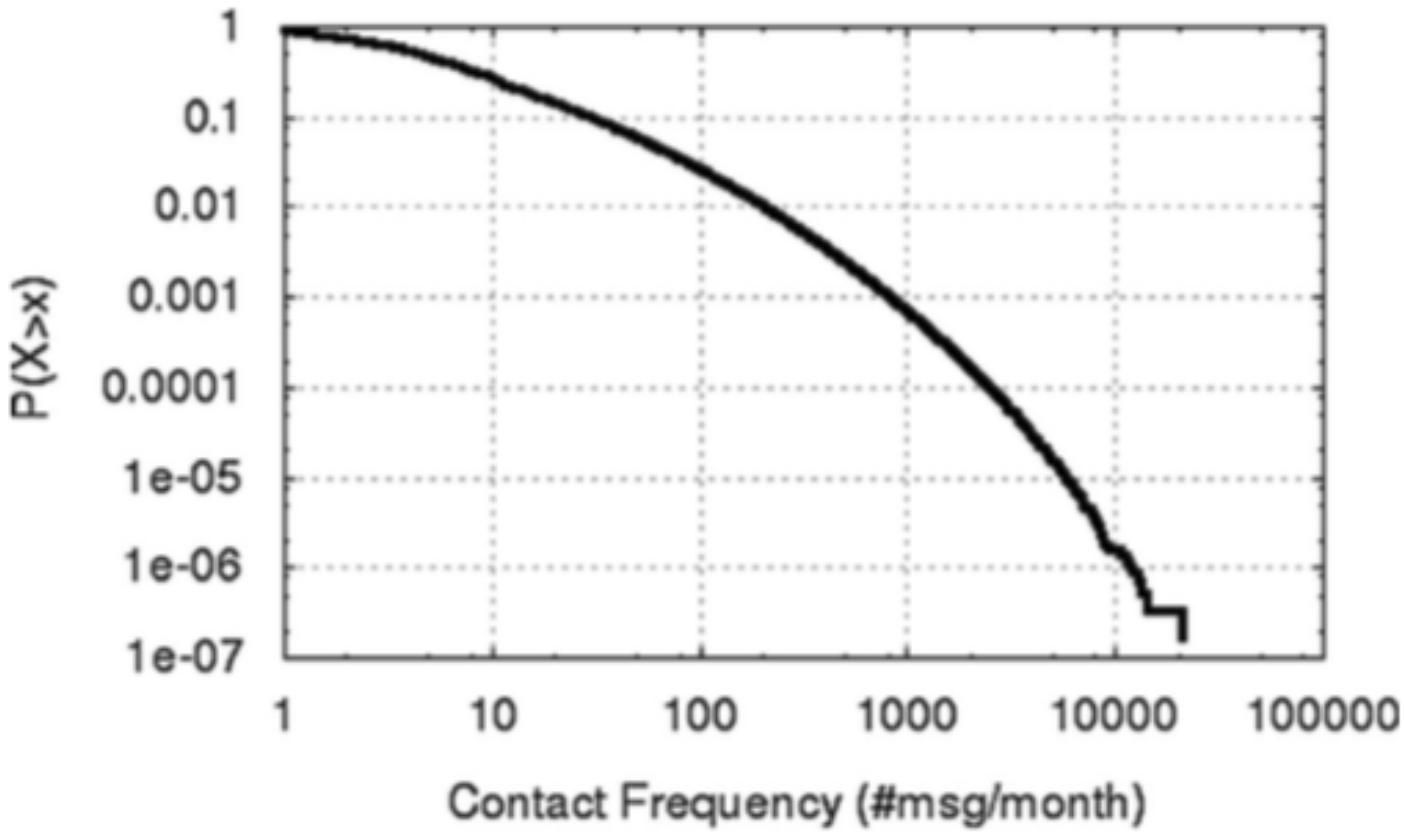

Fig. 7. CCDF of the contact frequency for relationships in Twitter.

Facebook contact info

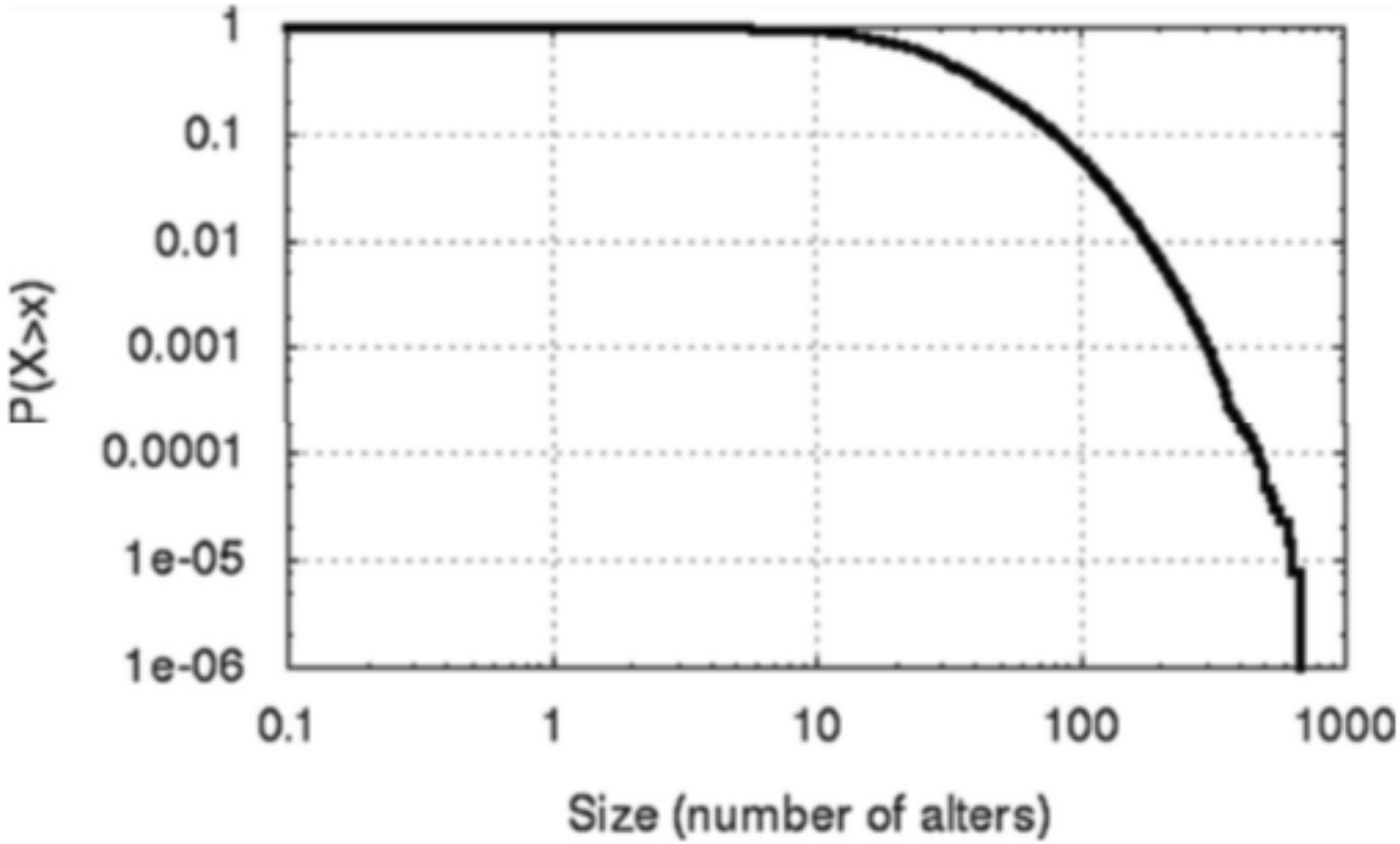

Fig. 8. CCDF of the size of ego networks in Twitter.

Facebook contact info

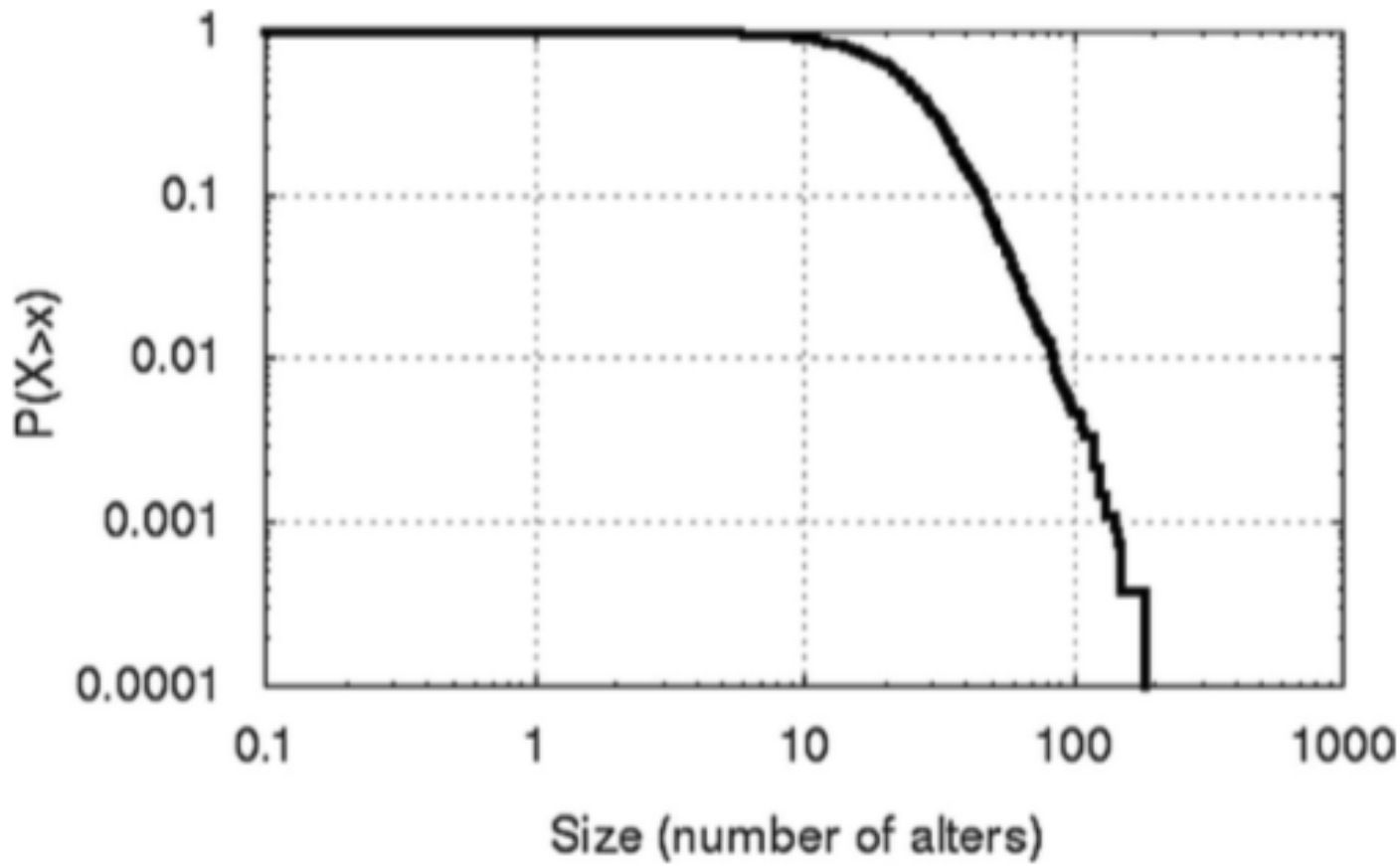

Fig. 9. CCDF of the size of ego networks considering only relationships with contact frequency higher than one message per year (active network) in Facebook dataset #2.

Analysis goal

- Cluster the frequency of contact of each ego network to search for layered structure.
- Tools: k-means and density based clustering
- Practice: for each ego, order the alters in a one dimensional space by contact frequency with the ego.
- Hard statistical part: how many clusters are there?
- Use penalization approach (AIC)

Penalization

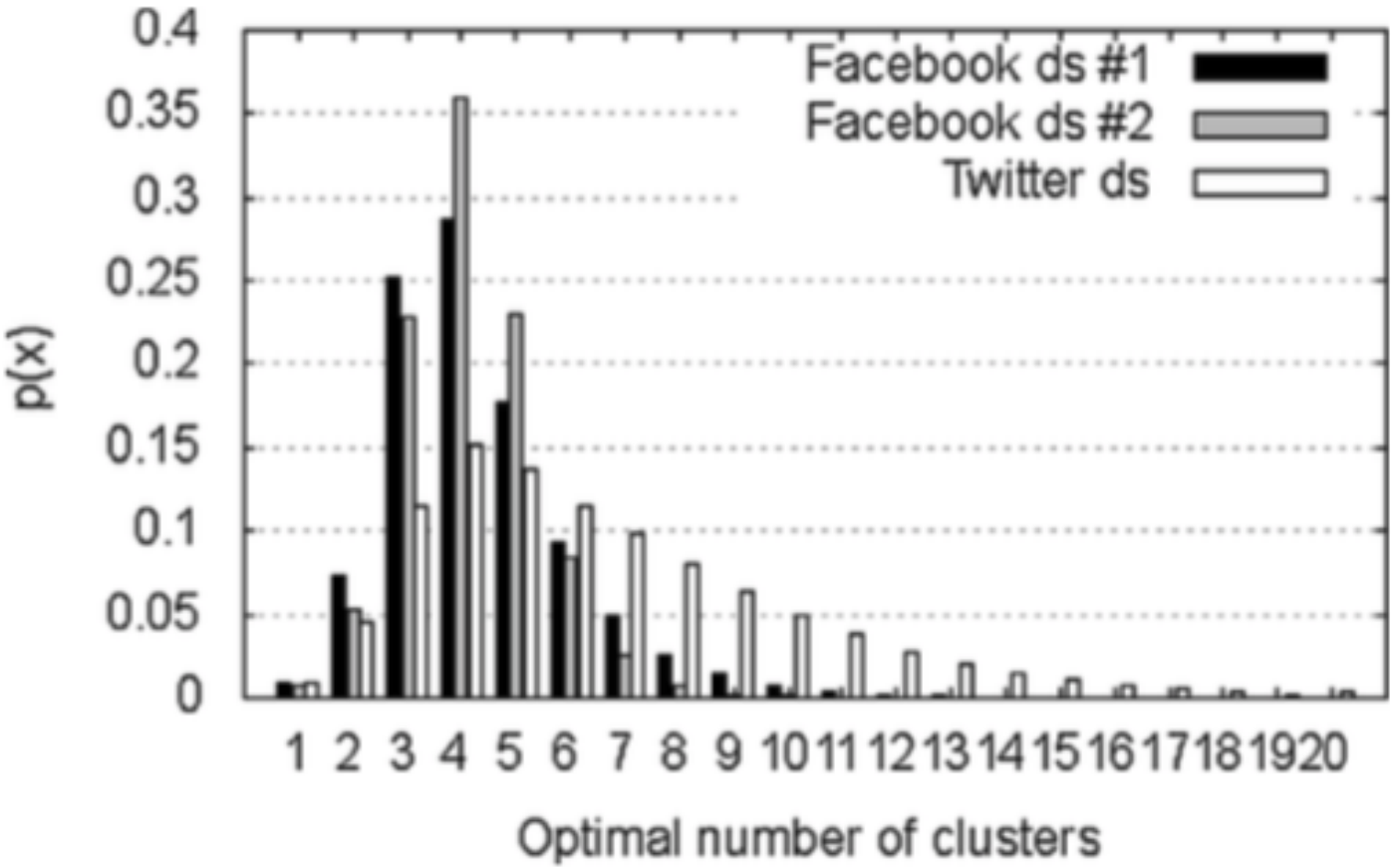

Fig. 10. Results of k-means cluster analysis for (a) Facebook dataset #1, (b) Facebook dataset #2 and (c) the Twitter dataset.

Find that 4 clusters for Facebook and 5 clusters for Twitter are "optimal"

Results

1. fFind evidence of conventional layer sizes (5,15 and 50) in both Twitter and Facebook.

Results

1. fFind evidence of conventional layer sizes (5,15 and 50) in both Twitter and Facebook.
2. fFind evidence of an outermost layer of size 150 for Twitter.

Results

1. fFind evidence of conventional layer sizes (5,15 and 50) in both Twitter and Facebook.
2. fFind evidence of an outermost layer of size 150 for Twitter.
3. fIdentify a "new layer that was not visible from face-to-face communication data" of size 1.5 individuals.

Results

1. fFind evidence of conventional layer sizes (5,15 and 50) in both Twitter and Facebook.
2. fFind evidence of an outermost layer of size 150 for Twitter.
3. fIdentify a "new layer that was not visible from face-to-face communication data" of size 1.5 individuals.
4. fCLaim: innermost layer has special relevance to egos due to high contact frequency.

Results

1. fFind evidence of conventional layer sizes (5,15 and 50) in both Twitter and Facebook.
2. fFind evidence of an outermost layer of size 150 for Twitter.
3. fIdentify a "new layer that was not visible from face-to-face communication data" of size 1.5 individuals.
4. fCLaim: innermost layer has special relevance to egos due to high contact frequency.
5. fConnection to "intimate friendship" literature where men have 0-1 friends and women have 1-2 friends.

Results

1. fFind evidence of conventional layer sizes (5,15 and 50) in both Twitter and Facebook.
2. fFind evidence of an outermost layer of size 150 for Twitter.
3. fIdentify a "new layer that was not visible from face-to-face communication data" of size 1.5 individuals.
4. fCLaim: innermost layer has special relevance to egos due to high contact frequency.
5. fConnection to "intimate friendship" literature where men have 0-1 friends and women have 1-2 friends.
6. fContact frequency is surprisingly similar to what's observed in face-to-face networks.