Network Metrology: Applications – Online Social Networks

This class

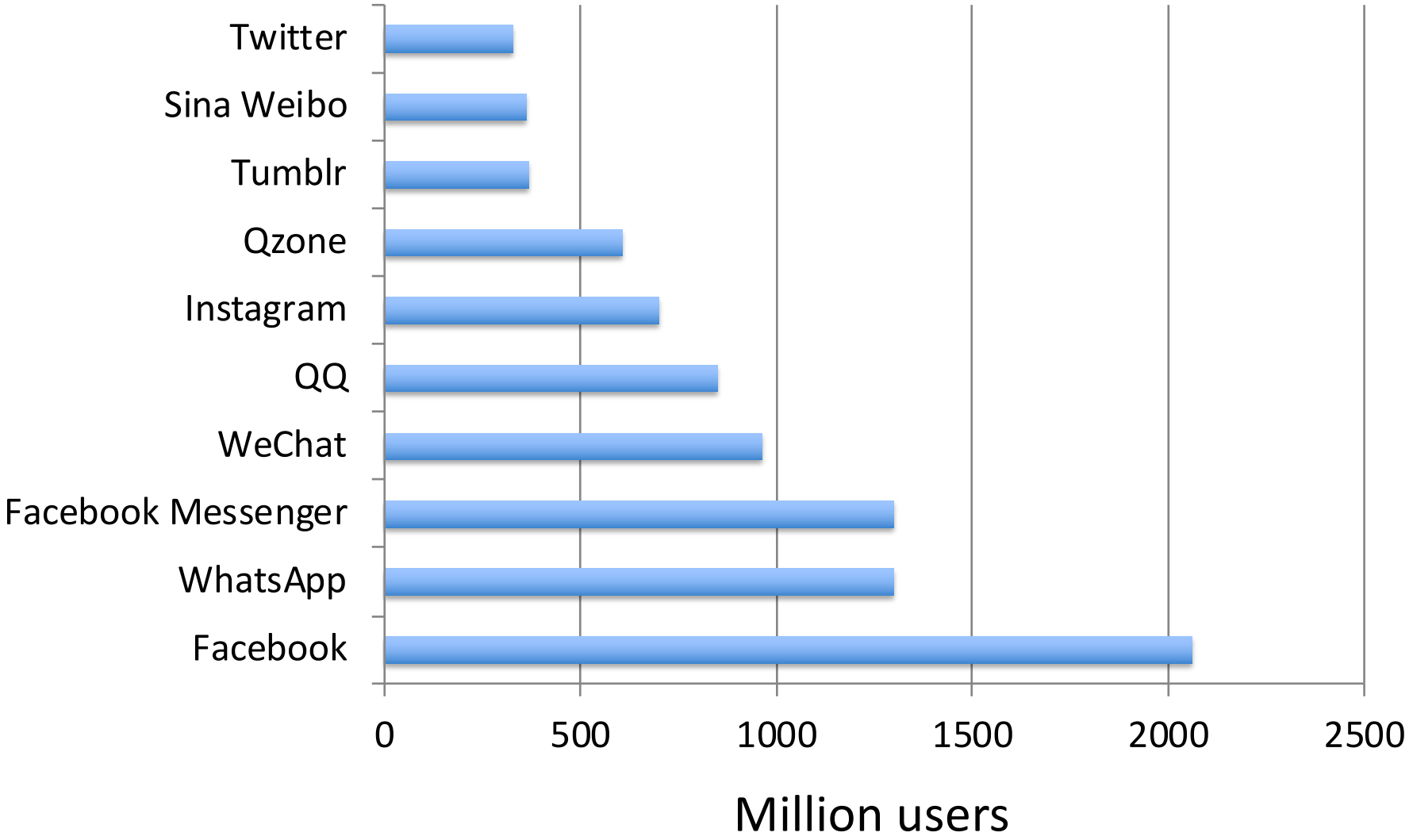
- Definition and motivation
  - What are online social networks?
  - Why measure them?
- Measurement methods
  - Crawling social networks
  - Sampling methods

DEFINITION & MOTIVATION

Online Social Networks (OSNs)

- Definition
  - Online system
    - Centered on users (who publish profiles)
    - Users create links to other users or content
    - Users can browse links and profiles
- Purpose
  - Maintain social ties
  - Upload/share content
  - Find new content

Most popular OSNs



http://www.statista.com/

Why measure OSNs?

- OSN developers
  - Content popularity/distribution
  - Trust relationships
- Advertisers, marketing specialists
  - Target ads to users based on profiles
  - Social influence in ads
- Sociologists, political scientists
  - Social dynamics
  - Social influence

How to measure OSN?

- OSN graph, G = (V,E)
  - V = users
  - E = relationship between users
- Evolution of OSN graph
  - How does G change over time?
- Content
  - What kind of information people share?

Basic characteristics

- Static properties of the graph
  - Number of users
  - Friend count distribution
  - Personal attributes (e.g., age, gender)
  - Sub-communities

Dynamic properties

- Capture temporal aspects and connectivity changes
  - Inter-communication frequency
  - Popularity growth
  - Rate of change of content

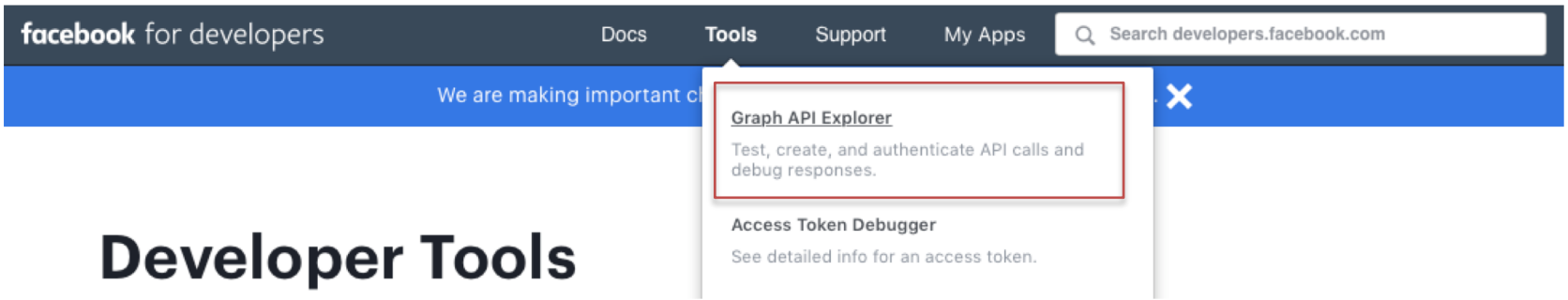
MEASUREMENT METHODS

Approaches

- Site operators
  - Direct access to databases of user profiles, links, posts
- Others
  - Crawling based on Web interface

Crawling the graph

- Basic method
  - Start with seed users, collect user profile
  - Follow links of friends, recursively request profiles
- Exact method will depend on OSN
  - HTML scraping
    - Requires getting information from the HTML page
    - Bandwidth intensive process
  - Leverage developer API



Question: How to select seed nodes?

- OSN graph is sparse
  - Seed nodes will bias the graph
- Solution: multiple seed nodes
  - If possible randomly selected

Challenges in crawling OSN

- OSN graphs are large and highly dynamic
  - Crawling takes time
- Some nodes are isolated
  - Can't reach through crawling
  - Focus on crawling weakly connected component
- OSNs limit access
  - Login requirements
  - Limited view
  - API query limits

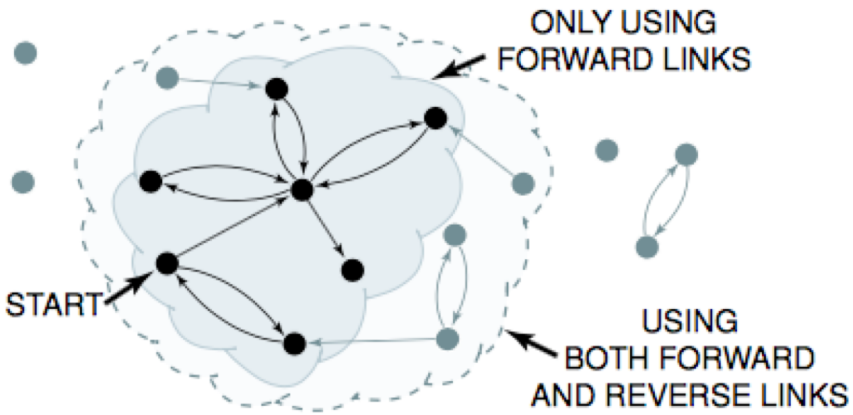


Fig 1. Users reached by crawling different link types

Source: Mislove, Alan, et al. "Measurement and analysis of online social networks." *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007.

Sampling methods

- Breadth first search (BFS)
  - Visit nodes in order of discovery
  - Incomplete BFS will cover just region of graph
- Random walk
  - Next node to visit selected uniformly at random among neighbors of current node
  - Biased towards high degree nodes
- Modified random walk to remove bias
  - Re-weighted or Metropolis-Hastings random walk
  - Estimates closer to random sampling

Considerations

- OSN term of use may forbid crawling
  - Must read terms before starting to crawl
- OSNs put limitations on API
  - Limit number of requests
- User's privacy settings may hide some or all information
  - Be aware that graphs can't capture all users

Summary

- Many reasons to measure OSNs
  - OSN developers, marketers, sociologists
- Crawling OSN is challenging
  - Large scale of networks
  - Limitations imposed by systems
- Method: sampling by graph crawling
  - Modified random walks better approximation of random sampling

References

- "Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems", A. Mislove
  - http://www.ccs.neu.edu/home/amislove/publications/SocialNetworks-Thesis.pdf
- "Practical Recommendations on Crawling Online Social Networks" JSAC 2011
  - http://mkurant.com/publications/papers/Gjoka_JSAC_2011_Practical.pdf