usfcs
UNIVERSITY of SAN FRANCISCO
department of computer science

De-anonymization

EJ Jung

11/01/10

usfCS
UNIVERSITY of SAN FRANCISCO
department of computer science

Anonymization techniques

- k-anonymity
  - mondrian
  - k-arq
- l-diversity
- t-closeness
- differential privacy

usfCS
UNIVERSITY of SAN FRANCISCO
department of computer science

De-anonymization techniques?

- Data in relational database
  - Linkage attack with auxiliary information
  - e.g. (gender, zip, birthday)
- Matrix data de-anonymization
  - Netflix dataset [NS08]
- Graph data de-anonymization
  - social graph de-anonymization [NS09]

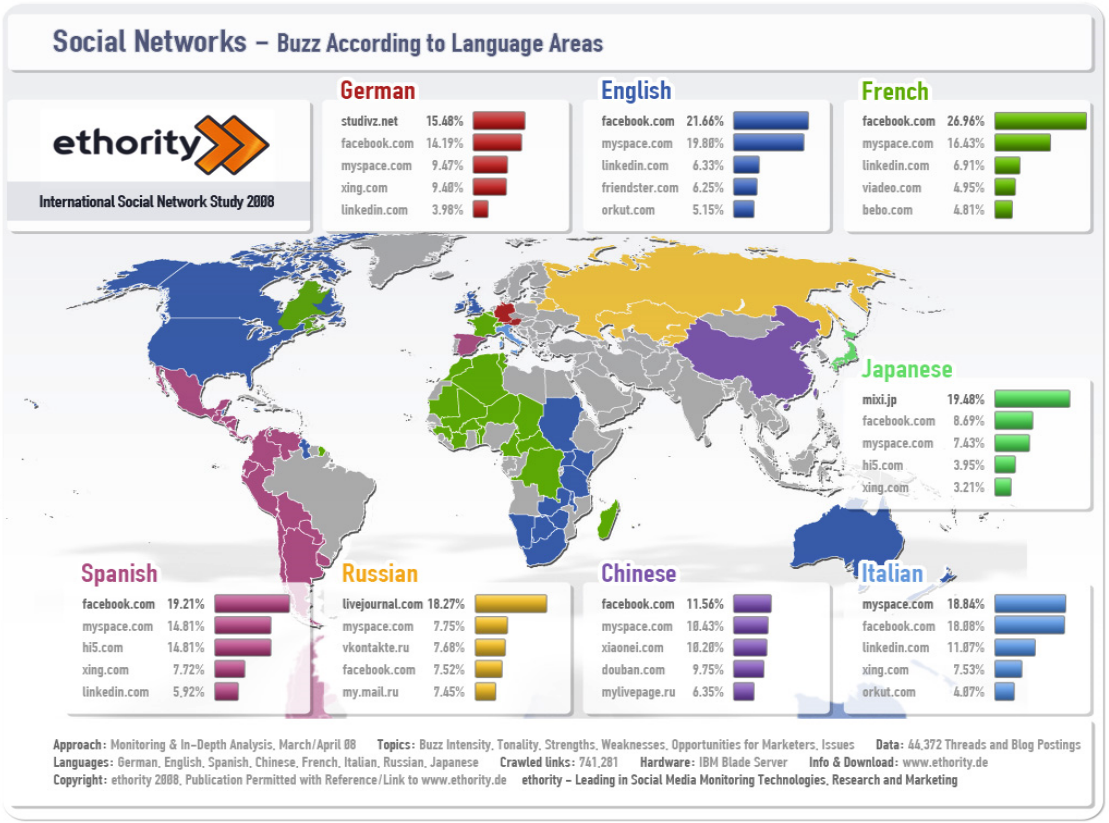
Social networks

- What is a social network?
  - an edge represents a social relationship
  - e.g. friendship, file download, email exchange, …
  - node is defined accordingly
- Note that overlay network is different
  - overlay network assumes "underlying" network
  - e.g. online social network is an overlay network over the Internet
- We only focus on online social network

usfCS
UNIVERSITY of SAN FRANCISCO
department of computer science

Examples of social networks

Social network websites



Small world (Watts&Strogatz, 1998)

- Most nodes are connected via a short path
- Small world graphs have a short diameter for a given size of |V| and |E|
- Social networks are small world networks.

Kevin Bacon's 6-degree law

History of social networks

- PGP network
  - edge = certificate
  - Alice vouches for Bob's public key
- Peer-to-peer network
  - edge = file sharing
  - Alice has downloaded (or uploaded) from Bob
- Social networking sites
  - edge = some form of friendship
  - Alice shares more about her with Bob
  - Facebook, myspace, livejournal, …

usfcs UNIVERSITY of SAN FRANCISCO department of computer science

Why do we care?

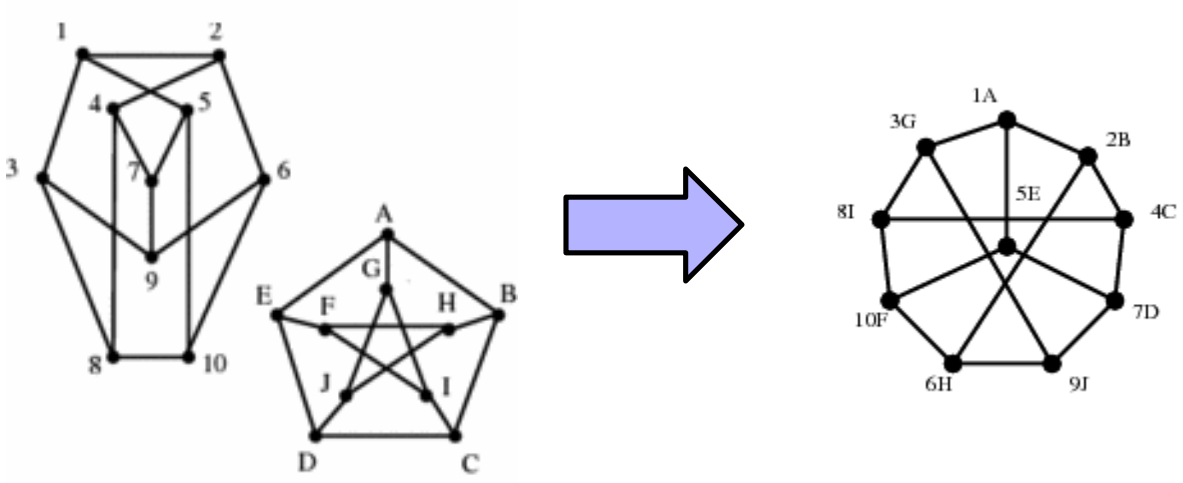
- Social network has pros
  - helps finding more users
  - helps identifying bad users
  - helps sharing reputation of users
    - builds its own supplement of "trust"
- Social network has cons
  - increases privacy and anonymity breach
  - serves as new attack vector
    - virus, worms, phishing, …

Notations

- Graph G=(V,E)
  - has a node set V and an edge set E.
  - n = |V|, e = |E|
  - <i, j> = edge from node i to j
  - we only do undirected edges today
- A path from node i to node j
  - <i, k1>,<k1,k2>,…,<kn,j>
  - this path has the length of (n+1)
- Diameter of a graph
  - the length of the longest path among
  - the shortest path between any pair of nodes

Graph isomorphism

- Input: two graphs G and H
- Output: mapping between nodes in G and H so that they are identical



- Autoisomorphism

http://www.cs.sunysb.edu/~algorith/files/graph-isomorphism.shtml

Subgraph isomorphism

- Input: two graphs G and G'
- Output: is G identical to a subgraph of G'?

- NP-Complete problem

- Applications
  - is this molecule part of a bigger molecule?
  - is this circuit part of another circuit?
  - is this social network appearing in the repository?

Cut the graph

- "Cut" of a graph is the sum of weights of the edges that are cut.
- Max cut, min cut, …

Walk the graph

Random walk brings to random places

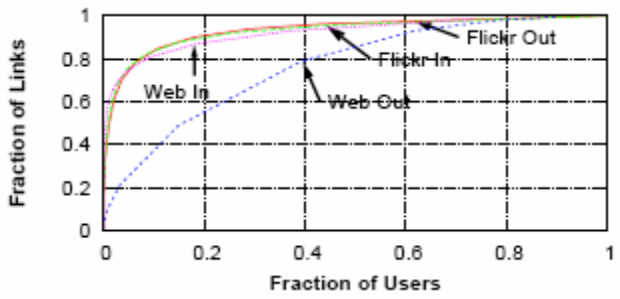
- Probability distribution $P_t(x)$ = prob. being at node x at time t.
- After a long walk, you may be anywhere.
  - $P_t(x) = P_{t+1}(x)$
- Stationary distribution
  - once you reach this distribution, $P_t(x) = P_{t+1}(x)$
  - how soon = mixing time

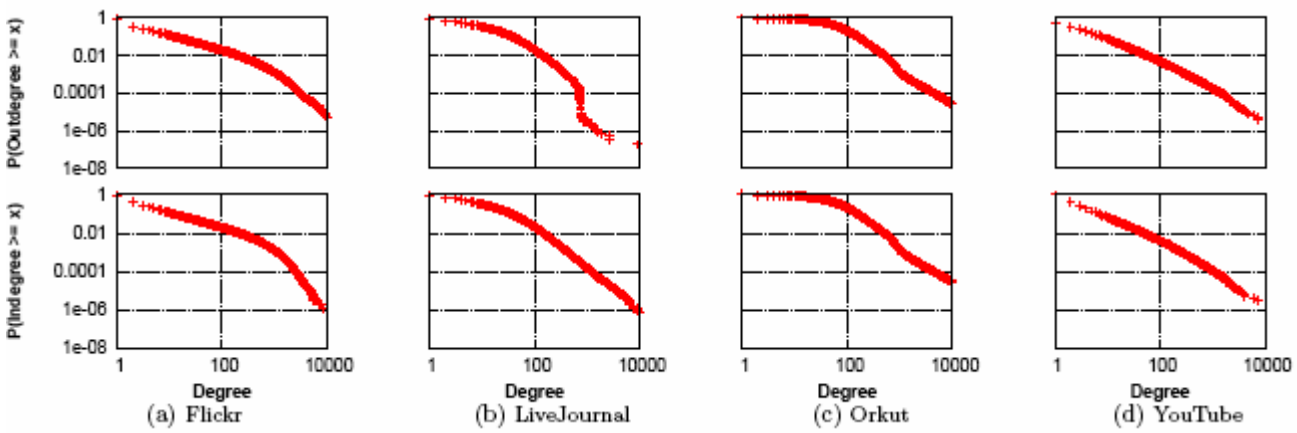
How random places?

- Note that Stationary distribution is $P_t(x) = P_{t+1}(x)$, not $P_t(x) = P_t(y)$
- Then what do we know about $P_t(x)$ and $P_t(y)$?
  - A connected non-bipartite undirected graph has a stationary distribution proportional to the degree distribution
  - More friends, more likely to run into

Scale-free networks (Mislove et al, 2007)

- degree distribution follows power-law
  - power law function: $p(x) \propto L(x)x^{-\alpha}$
  - zipf distribution: $P_n \sim 1/n^a$





What can we do with this?

usfcs
UNIVERSITY of SAN FRANCISCO
department of computer science

Big-O notation

Notation	Name	Intuition	As $n \to \infty$, eventually...	Definition
$f(n) \in O(g(n))$	Big Omicron; Big O; Big Oh	f is bounded above by g (up to constant factor) asymptotically	$f(n) \leq g(n) \cdot k$	$\exists(k > 0), n_0 : \forall(n > n_0)\ \lvert f(n)\rvert \leq \lvert g(n) \cdot k\rvert$ or $\exists(k > 0), n_0 : \forall(n > n_0)\ f(n) \leq g(n) \cdot k$
$f(n) \in \Omega(g(n))$	Big Omega	f is bounded below by g (up to constant factor) asymptotically	$f(n) \geq g(n) \cdot k$	$\exists(k > 0), n_0 : \forall(n > n_0)\ \lvert g(n) \cdot k\rvert \leq \lvert f(n)\rvert$
$f(n) \in \Theta(g(n))$	Big Theta	f is bounded both above and below by g asymptotically	$g(n) \cdot k_1 < f(n) < g(n) \cdot k_2$	$\exists(k_1, k_2 > 0), n_0 : \forall(n > n_0)\ \lvert g(n) \cdot k_1\rvert < \lvert f(n)\rvert < \lvert g(n) \cdot k_2\rvert$

http://en.wikipedia.org/wiki/Big_O_notation

Bayes' Theorem

- Conditional probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- Bayes' Theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Bayesian in Spam [Graham 02, Sahami 98]

- Given a word "stock" in an email, what is the probability of this email being spam?

we can compute these from the sample set of emails

$$P(spam \mid stock) = \frac{P(stock \mid spam)P(spam)}{P(stock)}$$

usfcs
UNIVERSITY of SAN FRANCISCO
department of computer science

De-anonymization

- Author's PPT on netflix de-anonymization
- De-anonymization overall by Narayanan
- Netflix lawsuit

De-anonymizing social networks