Users Joining Multiple Sites: Distributions and Patterns

Reza Zafarani and Huan Liu
Computer Science and Engineering, Arizona State University, USA
{Reza, HuanLiu} @asu.edu

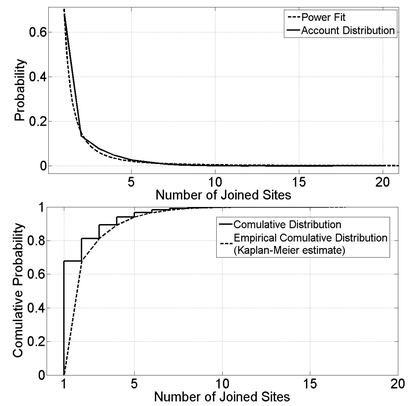
Users Join Multiple Sites!

- Our social media life is no longer limited to a single site. We post on Reddit, like on Facebook, tweet on Twitter, watch on YouTube, listen on Pandora, among many other activities exhibited by social media users.
- Users prefer more engaging sites, where they can find familiar faces such as friends, relatives, or colleagues.
- On average, popular sites with more members are expected to contain more friends for an average individual.

Question:
Does this fully explain users' site selections?

User Membership Distribution across Sites

- More than 97% of users have joined at 1 to 5 sites.
- A power function ($g(x) = 0.6761x^{-2.157}$) with 95% confidence fits to the curve with $R^2 = 0.9978$
- A maximum likelihood estimation shows that the distribution is *Power-Law*

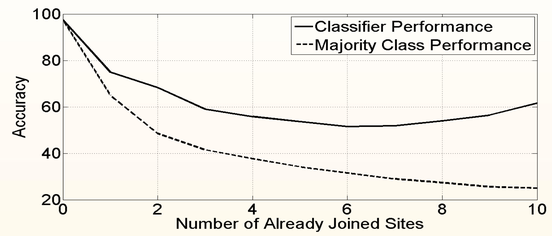


Evaluating via Recommending Sites to Users

- By identifying the type of site selection patterns a user has exhibited in the past, we recommend new sites to the user
- For users that have joined *n* sites:
  - We assume that given the category of *n-1* of them, the category of the *n*th site should be predictable.
- We generate all the possible combinations of n-1 sites and use the number of sites in each category as features (4 features) and the category of the *n*th site as the label.

Technique	AUC	Accuracy
J48 Decision Tree Learning	0.880	**79.25%**
Random Forest	0.895	79.17%
Logistic Regression	0.886	79.14%
SMO (Sequential Minimal Optimization)	0.728	78.92%
Naive Bayes	0.869	76.66%

- When users haven't joined any sites, they join popular sites: majority prediction is as accurate.
- As users join more sites, preference play an important role: majority prediction = random.

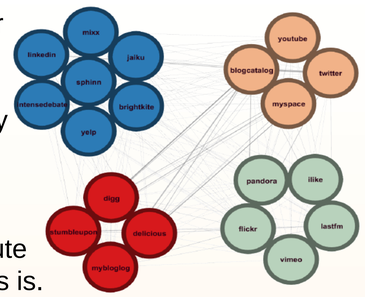


Data Preparation

- We can survey individuals for their accounts
  - Not Scalable + Expensive
- We can also utilize automatic approaches to connect corresponding identities of users across sites.
- Users often list their accounts on social networking sites, blogging and blog advertisement portals, and forums.
- We collected 96,194 users having accounts on a subset of 20 social media sites:
  - *BlogCatalog, BrightKite, Del.icio.us, Digg, Flickr, iLike, IntenseDebate, Jaiku, Last.fm, LinkedIn, Mixx, MySpace, MyBlogLog, Pandora, Sphinn, StumbleUpon, Twitter, Yelp, YouTube, and Vimeo*

User Membership Patterns across Sites

- We find sites that users join together
- If users join sites with a probability that is proportional to their popularity
  - The expected overlap between two sites is $\frac{d_i d_j}{2m}$.
  - Given the actual overlap between the two sites, O_{ij} , we can compute how non-random joining both sites is.
  - The problem is reduced to weighted modularity.



- *There are sites that users join all to be able to access the content that becomes available on each one of them.*
- *There are popular sites that users join all (or most) to satisfy their basic needs (average user behavior).*
- *There are [unknown/new] sites that early adopters join.*

This work was supported, in part, by the ONR Research grants: N000141110527 and N000141410095.