Aggregating Content and Network Information to Curate Twitter User Lists

Derek Greene Barry Smyth
Gavin Sheridan Pádraig Cunningham

RecSys-12: Recommender Systems & The Social Web

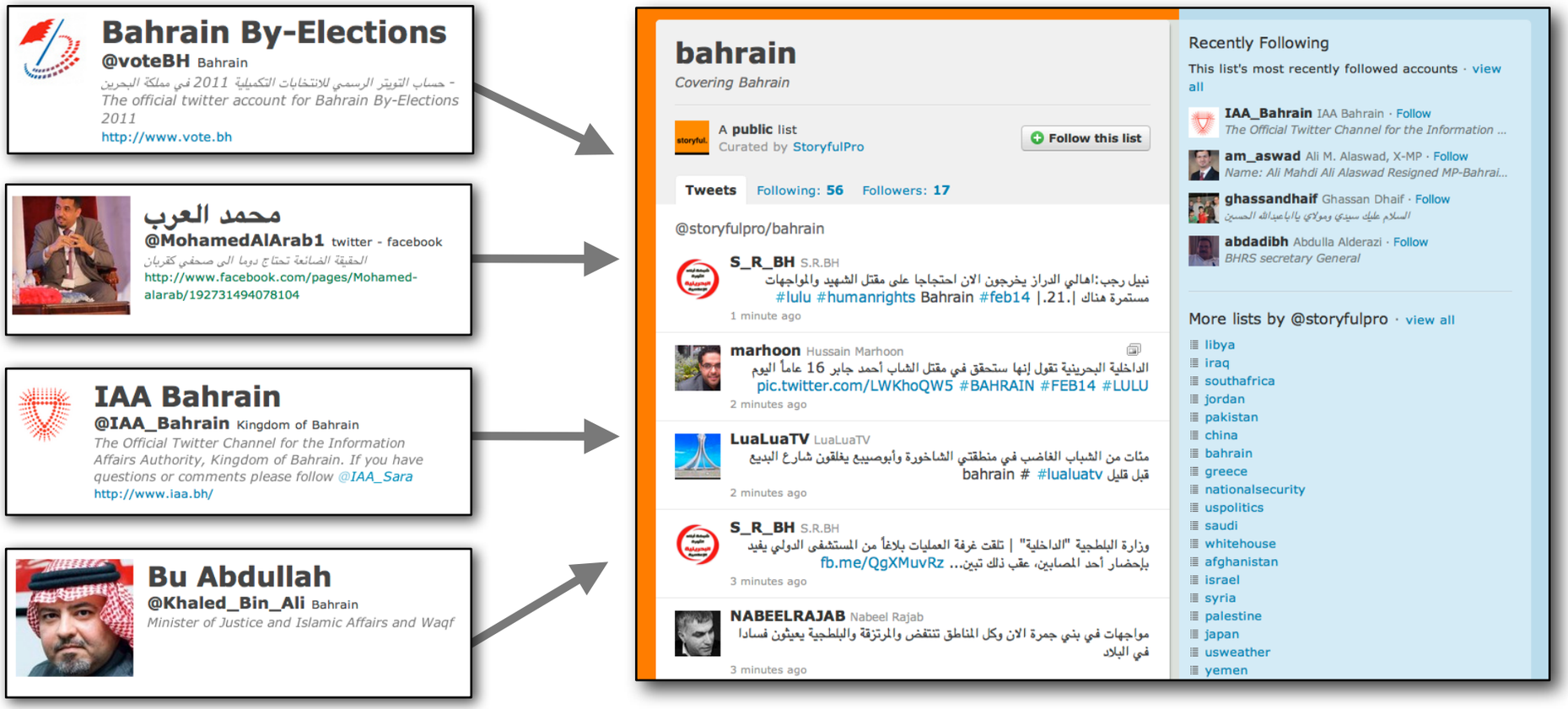
Clique: Graph & Network Analysis Cluster
School of Computer Science & Informatics
University College Dublin, Ireland

Task: User List Curation

- Twitter allows the grouping of users into topical **user lists**.
- **Storyful** maintains lists of users for news stories to monitor breaking news related to that story.

Supporting User List Curation

- **Idea:** Use data analysis to identify important users that form the "community" around a news story on Twitter.
- **Goal:** Recommend new users to expand an embryonic seed list, helping to find valuable content relating to the story.



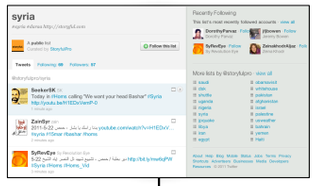
Curation System Overview

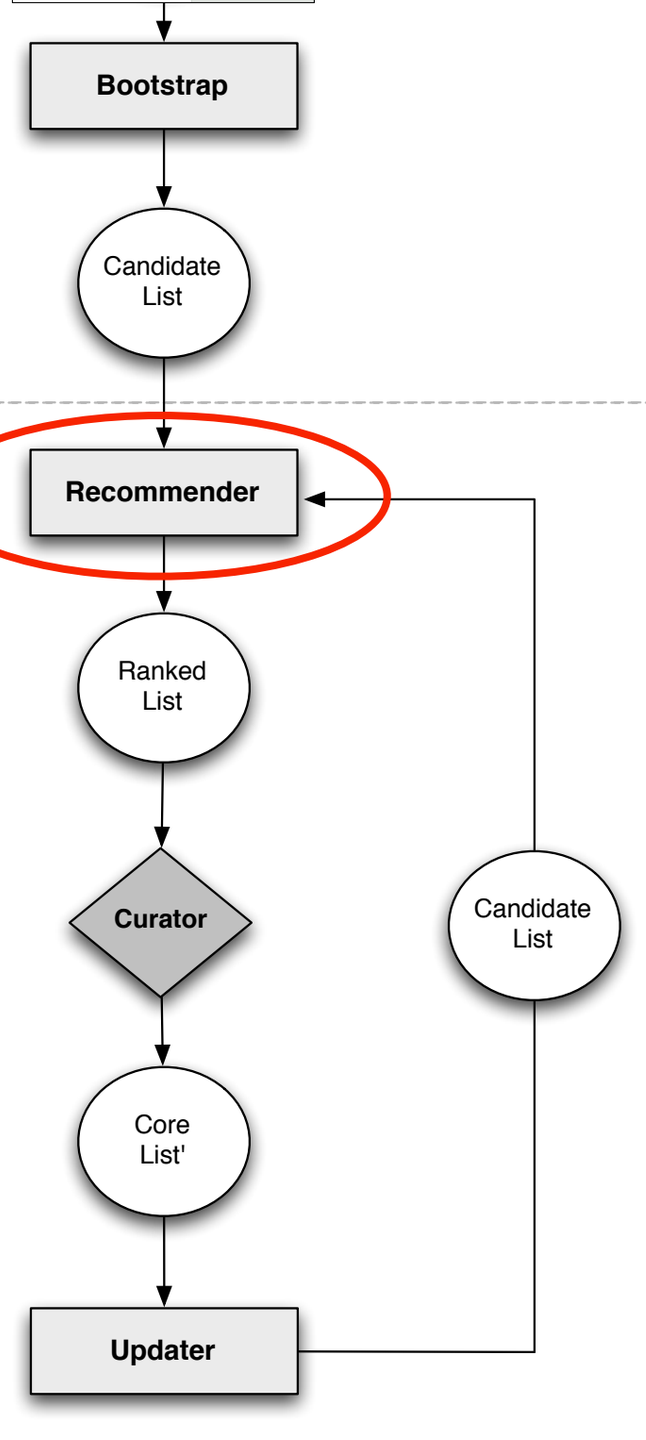
Clique

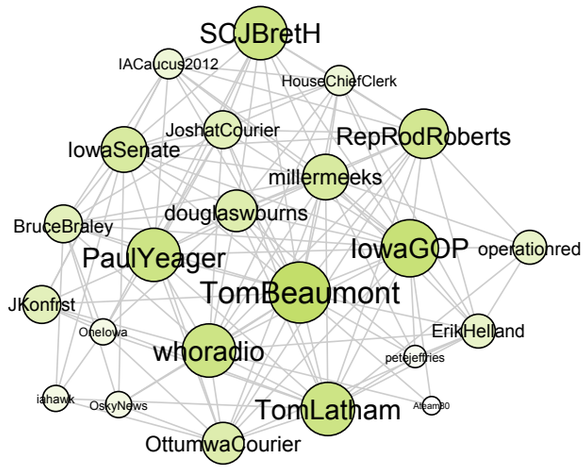
Seed Set

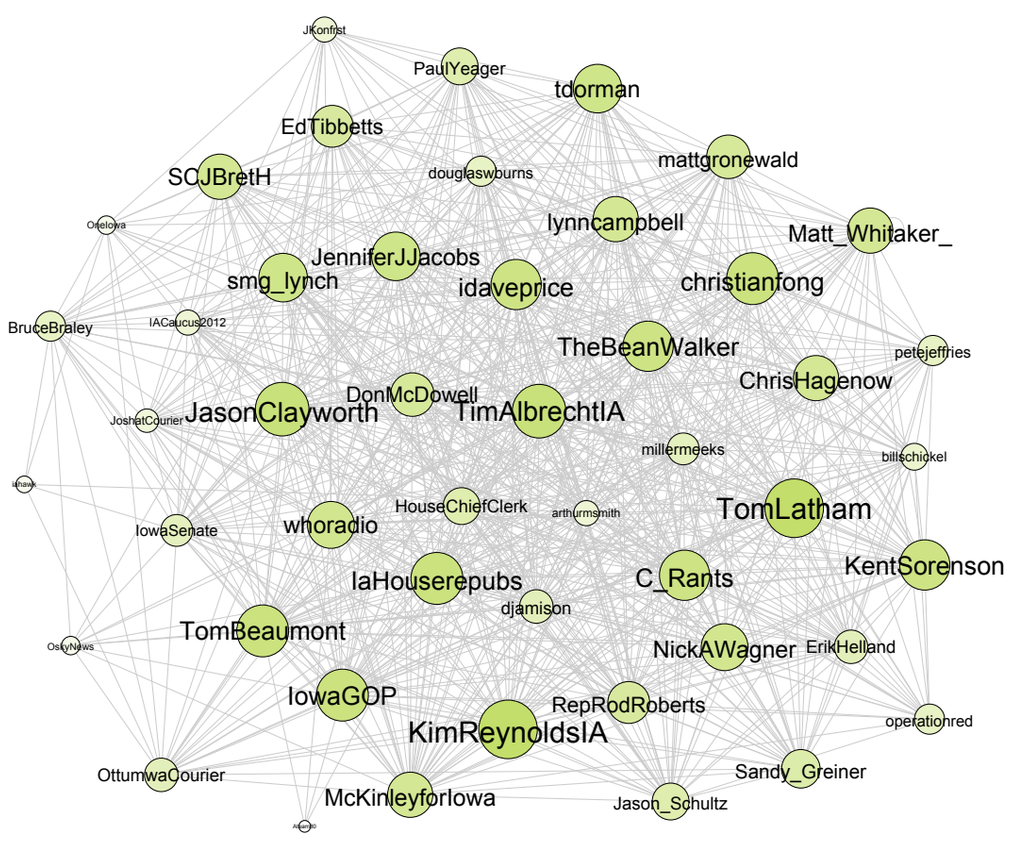
Phase 1

Phase 2







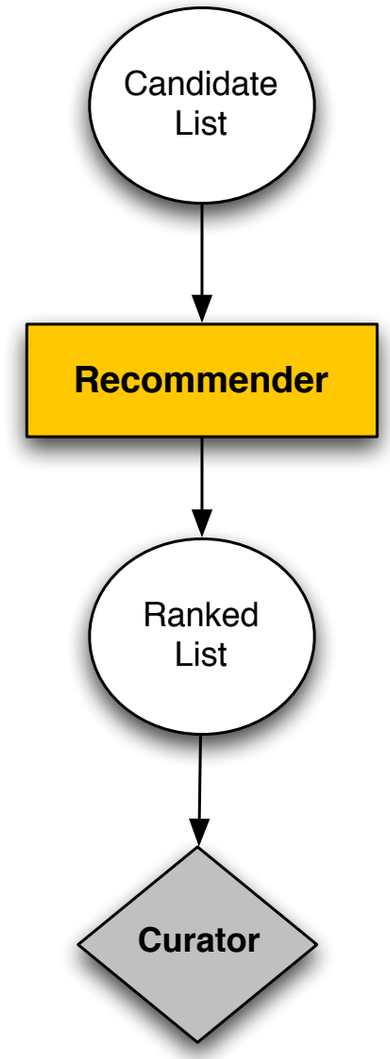


Recommendation Overview

General Recommender

- Use embryonic seed list as initial training data.
- Compute training data centroid.
- Measure cosine similarity between centroid and vectors representing candidate non-seed users.

➡ Produce ranked list of top *K* non-seed users to present to a human curator.



Which criterion should we use for recommendation?

➡ Propose variety of vectorised approaches, using criteria based on Twitter network and content analysis.

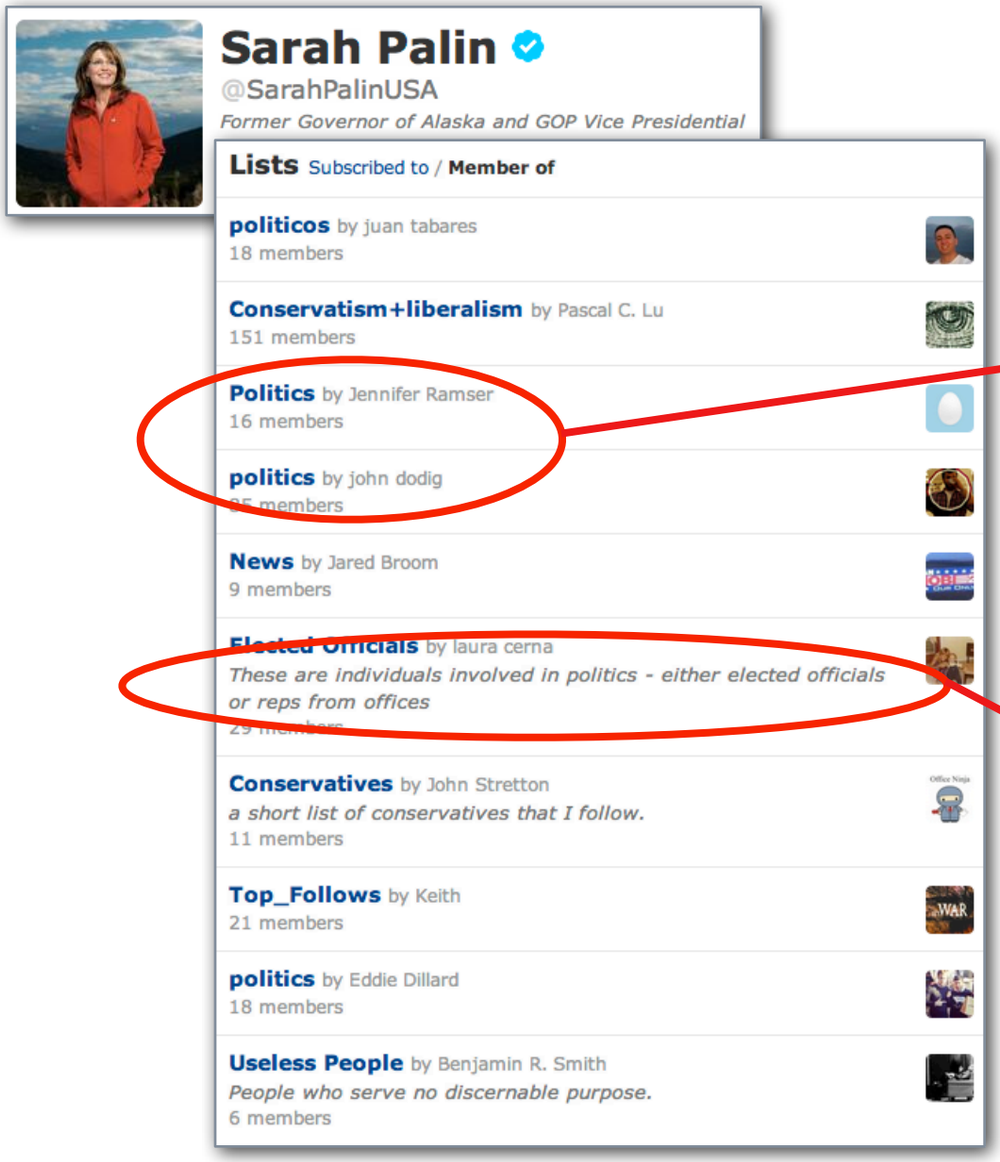
Content-Based Criteria

- **Tweet profiles:** Create a term vector for each user, containing the aggregation of their 50/100/200 most recent tweets.

User	**president**	**health**	**care**	**reform**	**gender**	**costs**	
@BarackObama	2	2	2	1	1	1	**...**

Content-Based Criteria

- **List names/descriptions:** Create a term vector for each user, containing the aggregation of names and/or descriptions of user lists to which they have been assigned.



User	**politics**	**news**	**useless**	
@SarahPalinUSA	2	1	1	...

User	**elected**	**politics**	**purpose**	
@SarahPalinUSA	1	1	1	...

Network-Based Criteria

- **Followed-by Profiles:** Users are similar if they are "co-followed" by the same set of users. Represent each user as a binary follower profile vector.

User X is followed by...

User X	**@TeamGB**	**@bradwiggins**	**@Mo_Farah**	**@britishswimming**
@MichaelPhelps	1	0	0	1
@chrishoy	1	1	1	0
@TomDaley1994	1	1	0	1

...

- **Mentioned-by Profiles:** Users are similar if they are mentioned in tweets by the same users (i.e. "co-mentioned)".

- **Retweeted-by Profiles:** Users are similar if their posts are retweeted by the same users (i.e. "co-retweeted)".

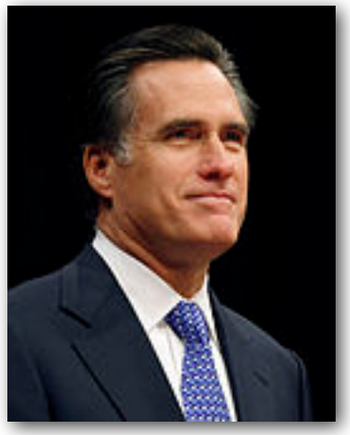
Network-Based Criteria

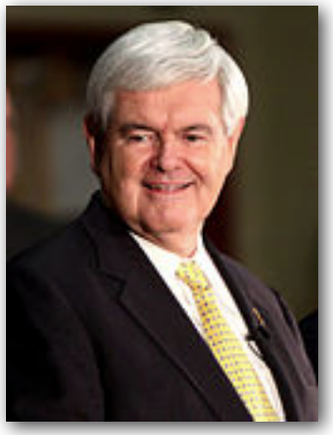
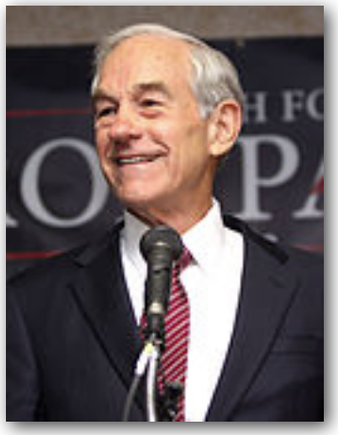
- **Co-Listed Information:**
  - Other media outlets and individuals will also be simultaneously curating user lists on topics in the wider Twitosphere.
  - ➡ Would like to "crowdsource" these efforts to support list curation.

Users @BrianODriscoll and @KearneyRob are co-listed on both lists.

Experimental Evaluation

- Collected 10 datasets around "Super Tuesday" GOP nominations in March 2012, with annotated seed users from Storyful.

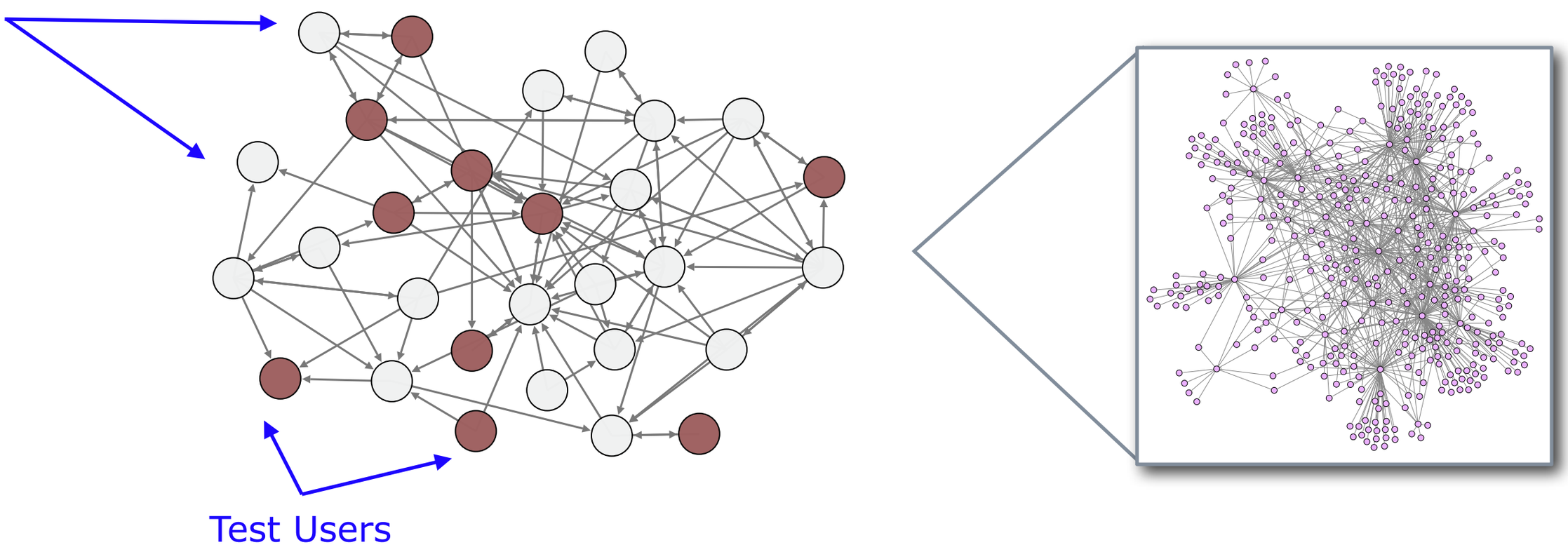


Dataset	Users	Core Users	Tweets	Friends	Followers	Listed
Alaska	948	41	185	208	269	89
Georgia	966	34	211	235	295	126
Idaho	743	20	186	264	273	47
Massachusetts	821	24	209	244	293	122
North Dakota	363	26	203	147	192	93
Ohio	1051	97	178	171	207	115
Oklahoma	693	32	205	178	211	109
Tennessee	979	48	199	170	204	112
Vermont	864	36	182	169	190	66
Virginia	877	46	200	160	199	115

Experimental Setup

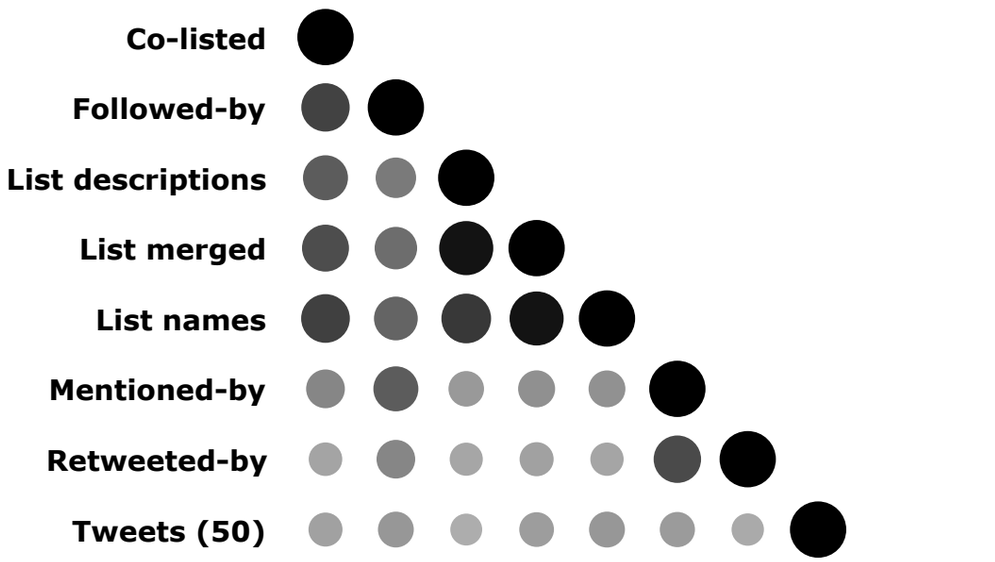
- Ran 250 X *k*-fold cross validation experiments per dataset:
  - Held out a proportion of the seed set as test data.
  - Used the remaining seed users as training data.
  - Ranked users in test data using each criterion.
  - Calculated *precision* and *recall* relative to the test data.

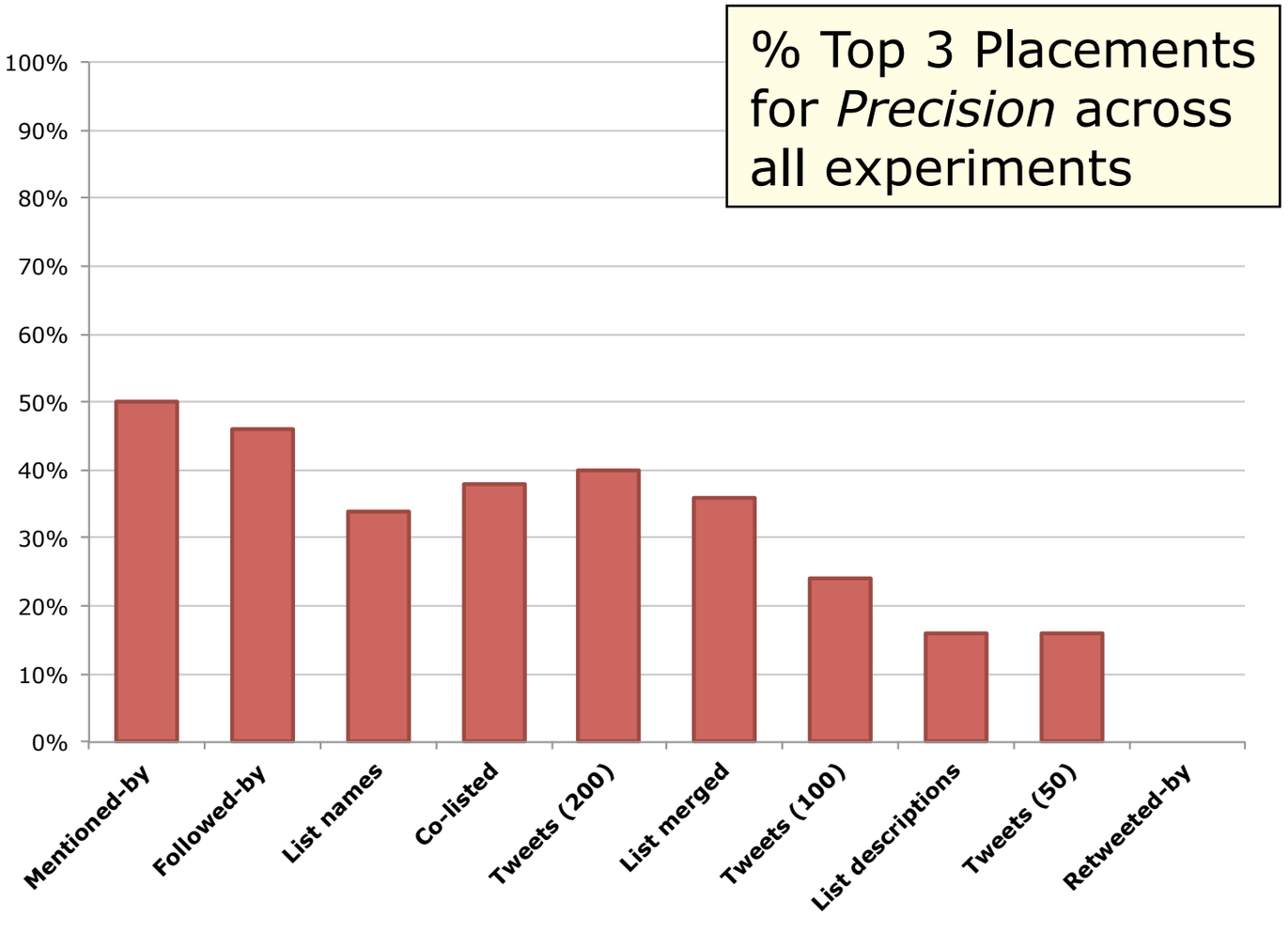


Comparison of Criteria

- Looked at diversity among the recommendations from the different criteria...

➡ We see several distinct signals present across the different views.

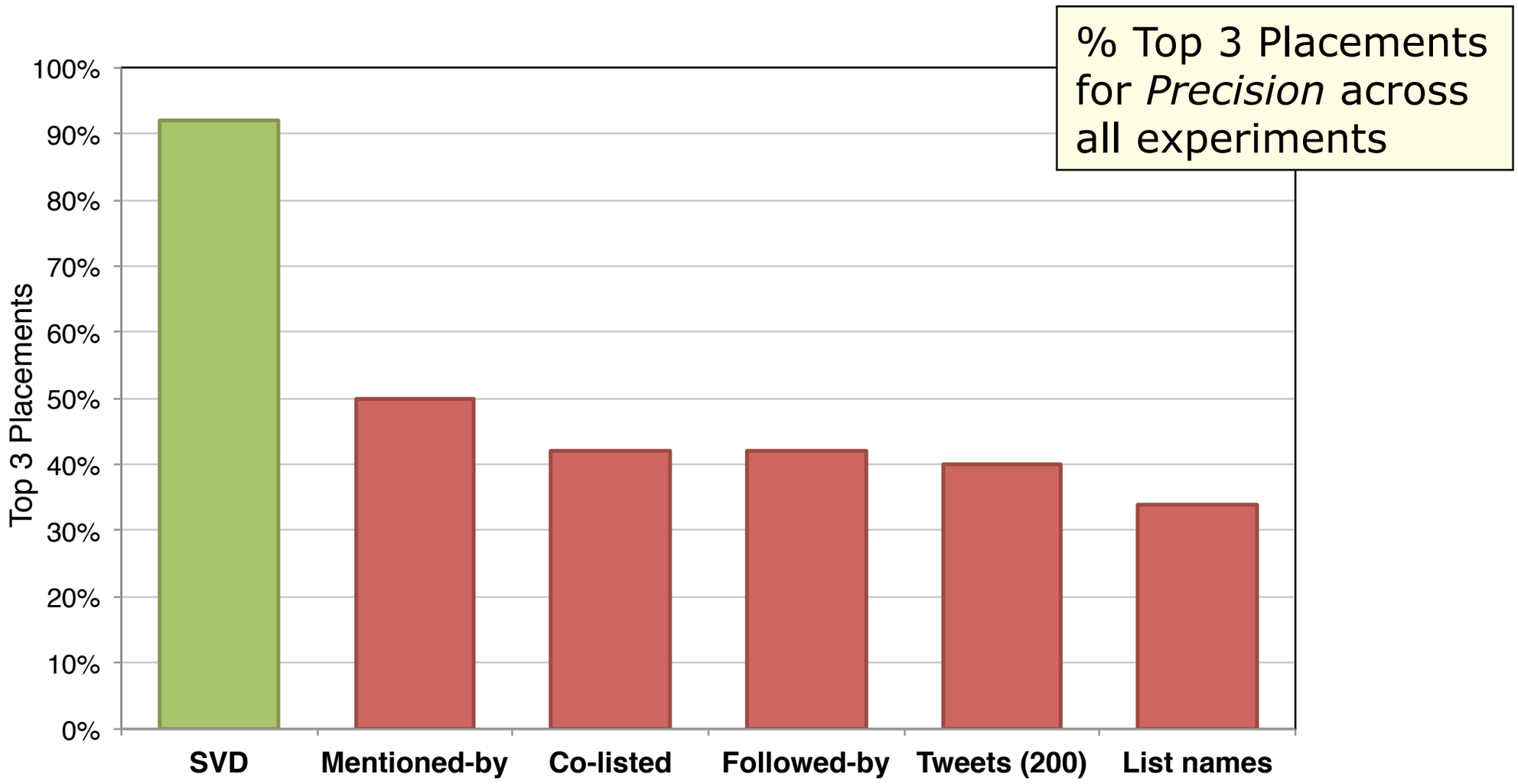




➡ No single criterion performs consistently well on all datasets.

Aggregating Multiple Rankings

- **SVD Aggregation:** Combine rankings generated on top 5 individual criteria into a single matrix, apply Singular Value Decomposition, then rank values in first singular vector.



Comparison of SVD Aggregation V Top 5 Individual Criteria

User List Curation System

Clique

storyful: Curation System

User Lists

Create New List...

Import List...

Manage Blacklists

>> afghanistan
>> africa
>> alabama
>> albania
>> algeria
>> andalusia
>> angola
>> argentina
>> arizona
>> armenia
>> art
>> asia-pacific
>> austria
>> azerbaijan
>> bahrain

Curation System Home

Recent Activity:

Last bootstrapped:	[greece] [portugal] [uk]
Last recommended:	[greece] [france] [it-security]
Last updated:	[greece] [france] [it-security]
Last modified:	[master-list]

Pending Instructions:

Project	Process Type	Requested At	Scheduled For
it-security	CONFIGURE	2012-06-22 12:08	2012-06-22 12:08
angola	UPDATE	2012-06-22 01:04	2012-06-22 12:16
afghanistan	RECOMMEND	2012-06-22 11:51	2012-06-22 12:24
portugal	UPDATE	2012-06-22 01:22	2012-06-22 12:35
france	UPDATE	2012-06-22 02:52	2012-06-22 13:22

Active Processes:

Project	Process Type	Started At	Status
armenia	BOOTSTRAP	2012-06-22 11:59	RUNNING
kosovo	BOOTSTRAP	2012-06-22 11:58	RUNNING
uk	UPDATE	2012-06-22 11:48	RUNNING

Conclusions

- **Summary:**
  - Proposed variety of criteria for user list building.
  - Performed a comprehensive comparison of individual criteria, demonstrating weaknesses of both content and network-based strategies.
  - Demonstrated that more accurate results can be achieved using SVD aggregation.

- **Future Work:**
  - Stratification of networks to target recommendations for niche communities - avoiding the "filter bubble".
  - Support for curation across multiple social networks.

Any Questions ?

derek.greene@ucd.ie

@derekgreene

http://cliquecluster.org