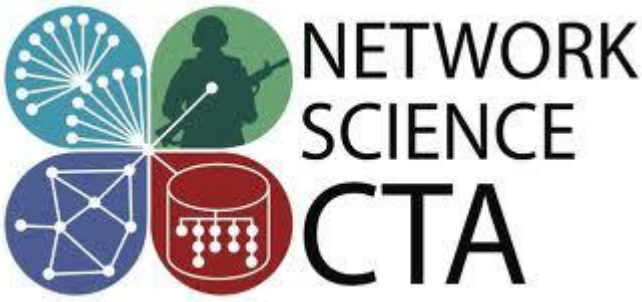
The Social Media Genome: Modeling Individual Topic-Specific Behavior in Social Media

Petko Bogdanov
Ambuj K. Singh
UCSB, Comp. Sci.

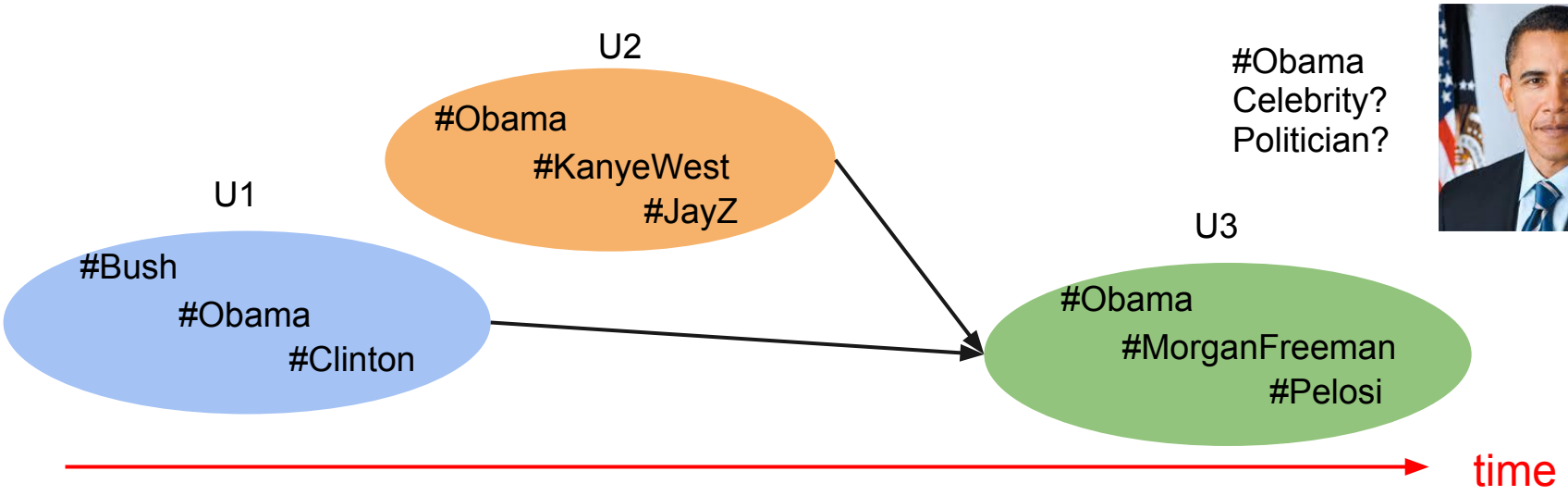
Michael Busch
Jeff Moehlis
UCSB, Mech. Eng.

Boleslaw K. Szymanski
RPI, Comp. Sci.

Michael Busch
ASONAM - Aug. 27, 2013

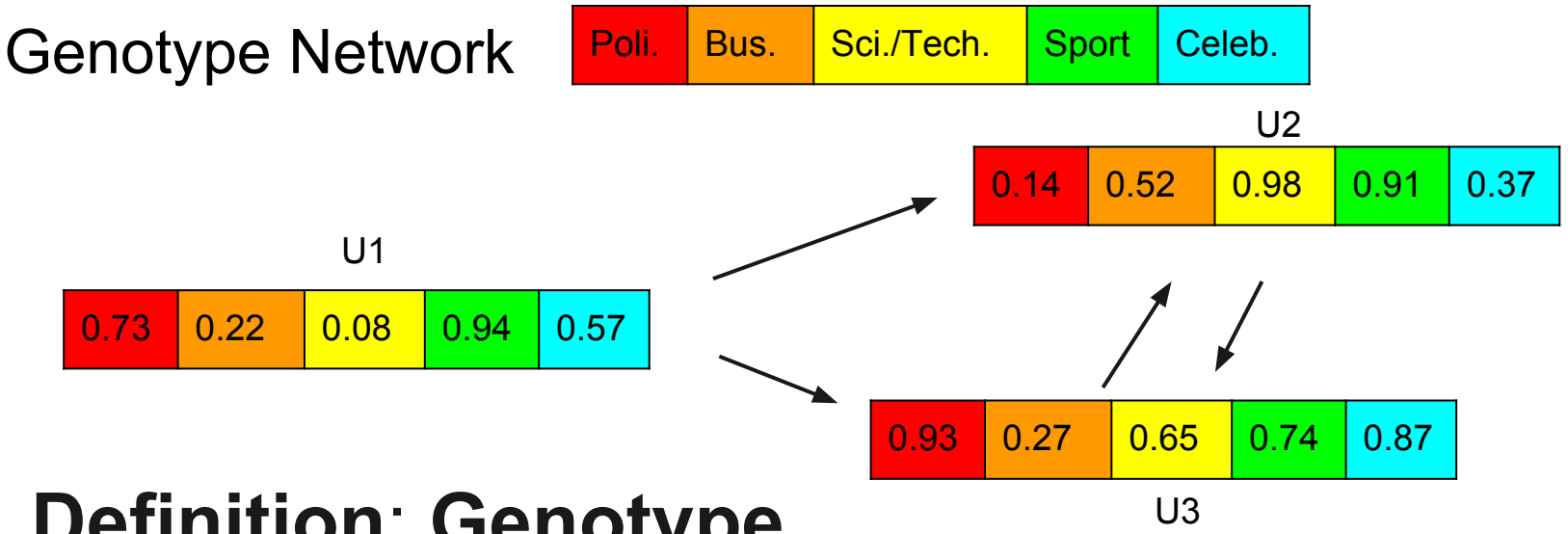


User Model: Motivation



- Predict influence on neighbors
- Dependent user behavior
  - Behavior adopted from neighbors (ex. Hashtags)
- Evolutionary descriptions (i.e., genetics)
  - How do behaviors change in time? (future work)

User Model: Genotype



Definition: **Genotype**

(1) a per-user entity that summarizes *observable behavior* of the user w.r.t different *topics*.

(2) an allele that the user introduces to the process of message propagation through a network.

Twitter Data

Proof of concept using Twitter messages containing hashtags:

- SNAP data set from 2009 (Leskovec)
- CRAWL data set from 2012 (our lab)

	SNAP (users=42M,tweets=467M)			CRAWL (users=9K,tweets=14.5M)		
Topic	Hashtags	Users	Uses/HT	Hashtags	Users	Uses/HT
Business	27	20k	1,155	19	1,493	88
Celebrities	32	26k	1,009	-	-	-
Politics	485	349k	2,020	121	5,480	49
Sci/Tech	33	415k	6,889	63	4,982	100
Sports	98	76k	3,274	24	320	14

TABLE : Statistics of the SNAP and CRAWL data sets.

Invariant behavior

Users are consistent in how they respond to a topic.

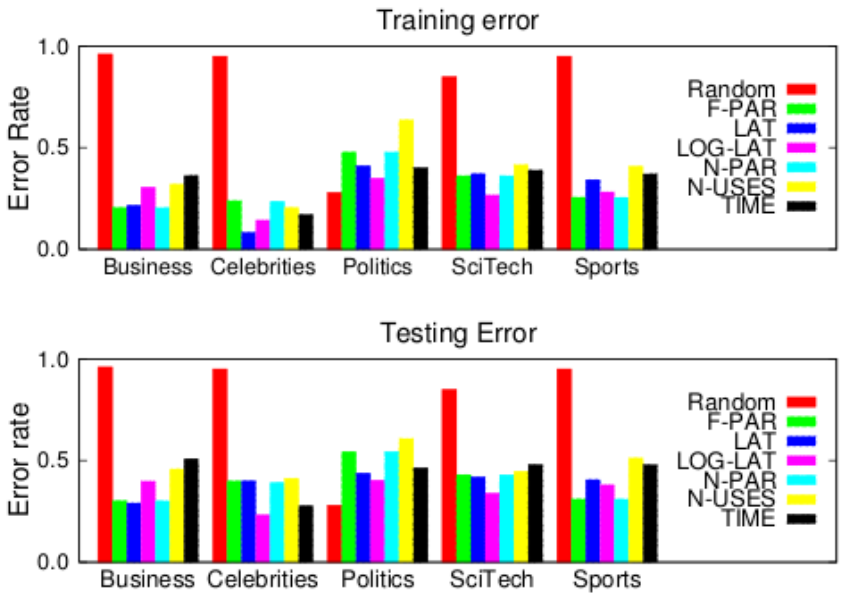


Fig. 1: Training and testing accuracy of leave-one-out Linear Discriminant (LD) classification.

HT metrics for each user:

Random	Randomly pick topic for HT. Proportional to prior distribution.
F-PAR	Fraction of parents who used HT.
LAT	Inverse of number of posts b/n first HT use of parent and user.
LOG-LAT	Log-normalized version of LAT.
N-USES	Number of HT uses.
TIME	Amount of time b/n first HT use of parent and user.

Network HT classifier

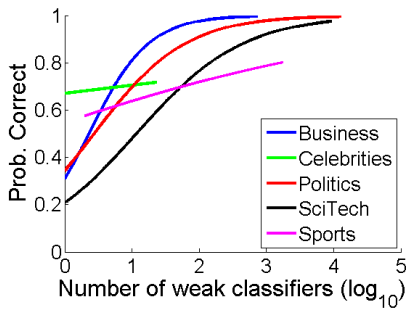
Observations:

(a) Accuracy improves with # of HT users

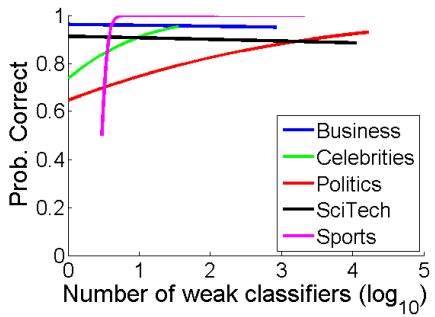
(b) LOG-LAT filters out avg. behavior, performs best

	Bus	Celeb	Pol	Sci./Tech	Sport	E[x]
Rand.	0.96	0.95	0.28	0.85	0.95	0.45
F-PAR	0.50	0.88	0.61	0.15	0.09	0.41
LAT	0.09	0.46	0.18	0.19	0.25	0.21
LOG-LAT	0.05	0.13	0.19	0.12	0.03	0.13
N-PAR	0.09	0.50	0.88	0.09	0.03	0.40
N-USES	0.45	0.42	0.90	0.22	0.56	0.54
TIME	1.0	1.0	0.01	0.92	0.88	0.61

Table III: Error rates of the Naive-Bayes (NB) consensus topic classification.



(a) LAT Net Classifier



(b) LOG-LAT Net Classifier

Fig. 2: Accuracy of the network classification as a function of the number of weak (local) classifiers.

Influence Backbones

- ***Influence Edge*** = directed edge connecting a user of a HT to all of his followers who use the same HT at a later time.
- ***Influence Network*** (Backbone) = a subset of the follower network, made of influence edges.
  - When sorted by topic: influence edge weights are proportional to number of HT influence edges of same topic.

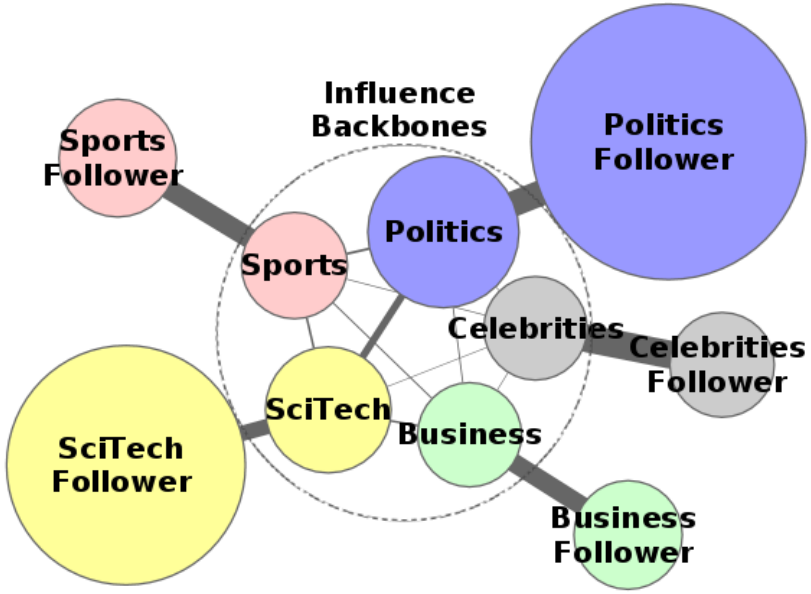


Fig. 3: Overlap among topic influence and corresponding follower subnetworks (SNAP).

Influence Backbones

- ***Influence Edge*** = directed edge connecting a user of a HT to all of his followers who use the same HT at a later time.
- ***Influence Network*** (Backbone) = a subset of the follower network, made of influence edges.
  - When sorted by topic: influence edge weights are proportional to number of HT influence edges of same topic.
- Relatively small SCC of influence networks supports the existence of influential *root* nodes.
- Kendall-tau rank correlations show that backbone rank is dissimilar to other ranks.

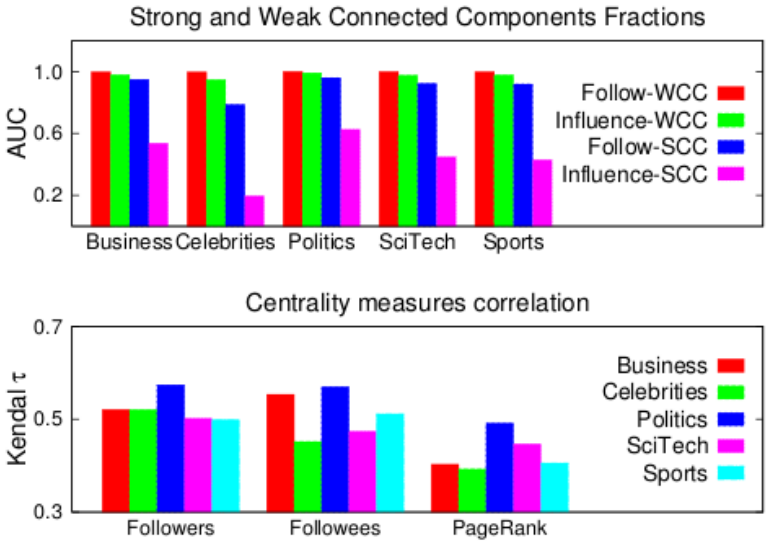


Fig. 4:
(Top) Largest weakly and strongly connected component sizes as a fraction of the network size.
(Bottom) Kendall-tau rank correlation of node importance measures between influence and follower networks.

Application: Influence Prediction

★ Activity-based (genotype) predictors perform 20% better than structural predictors.

★ Genotype + Backbone structure outperforms all others.

- Structural Predictors:
  - # of Followees
  - # of Followers
  - # of reciprocal links
- Activity-based predictors:
  - Act = same HT history
  - Topic Act = same topic history
  - RW+Act = Backbone centrality + Act

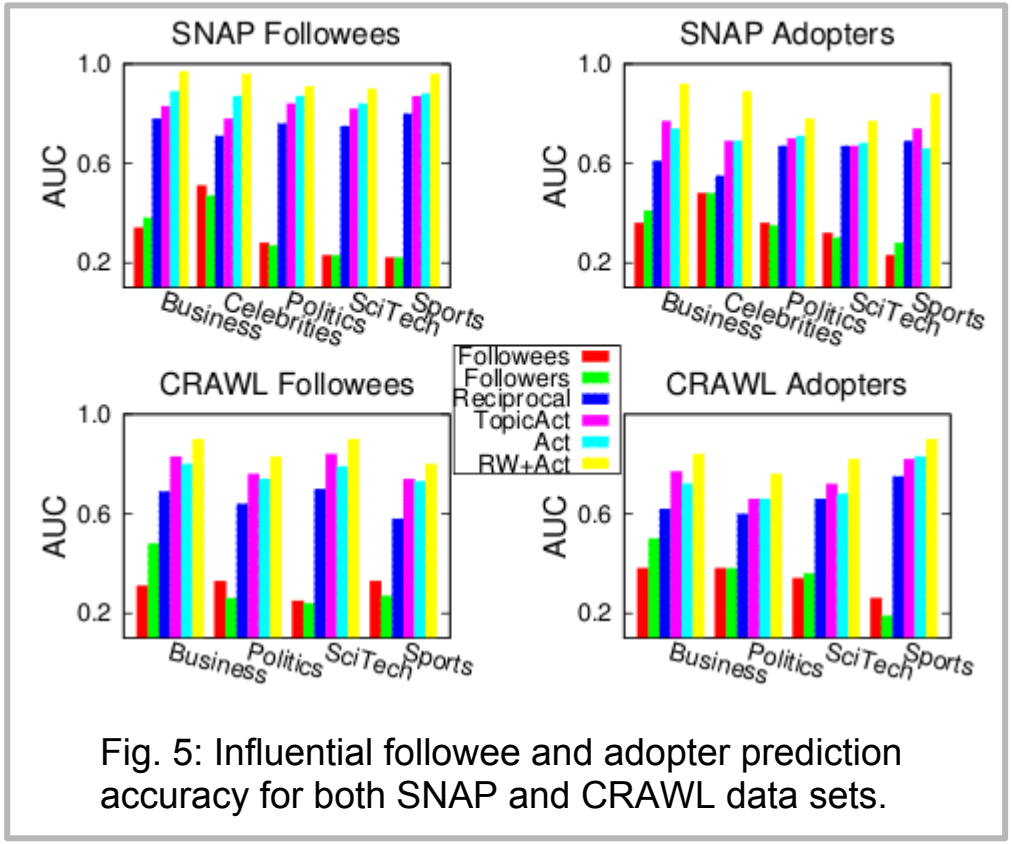


Fig. 5: Influential followee and adopter prediction accuracy for both SNAP and CRAWL data sets.

Application: Network Latency Minimization

- ★ 40% latency reduction by targeting 1% of the network.
- ★ Latency minimization requires both genotype and influence backbone.

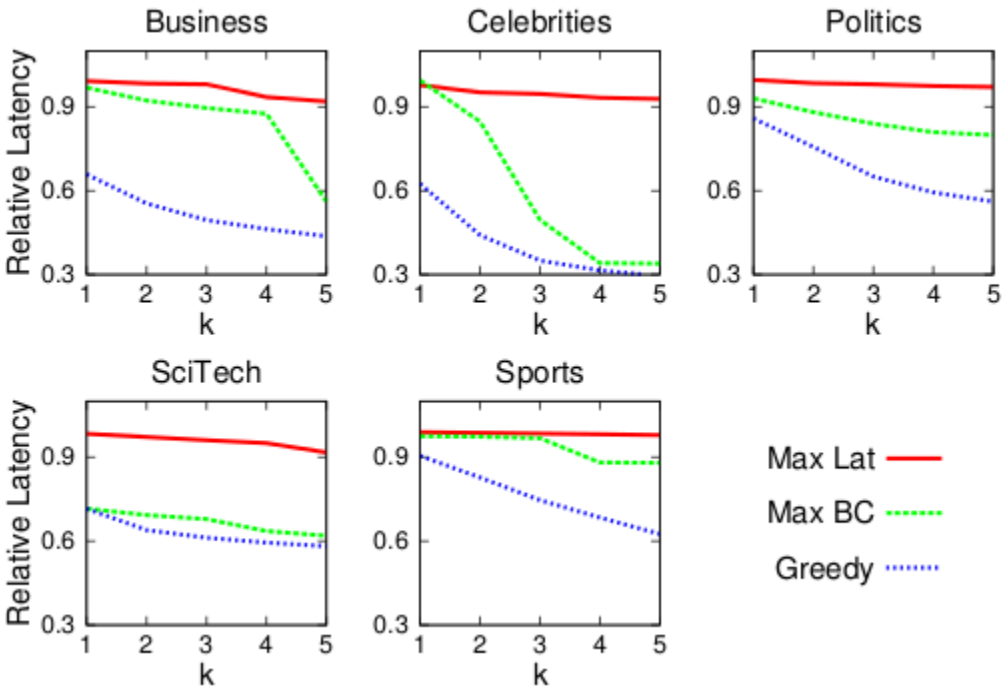
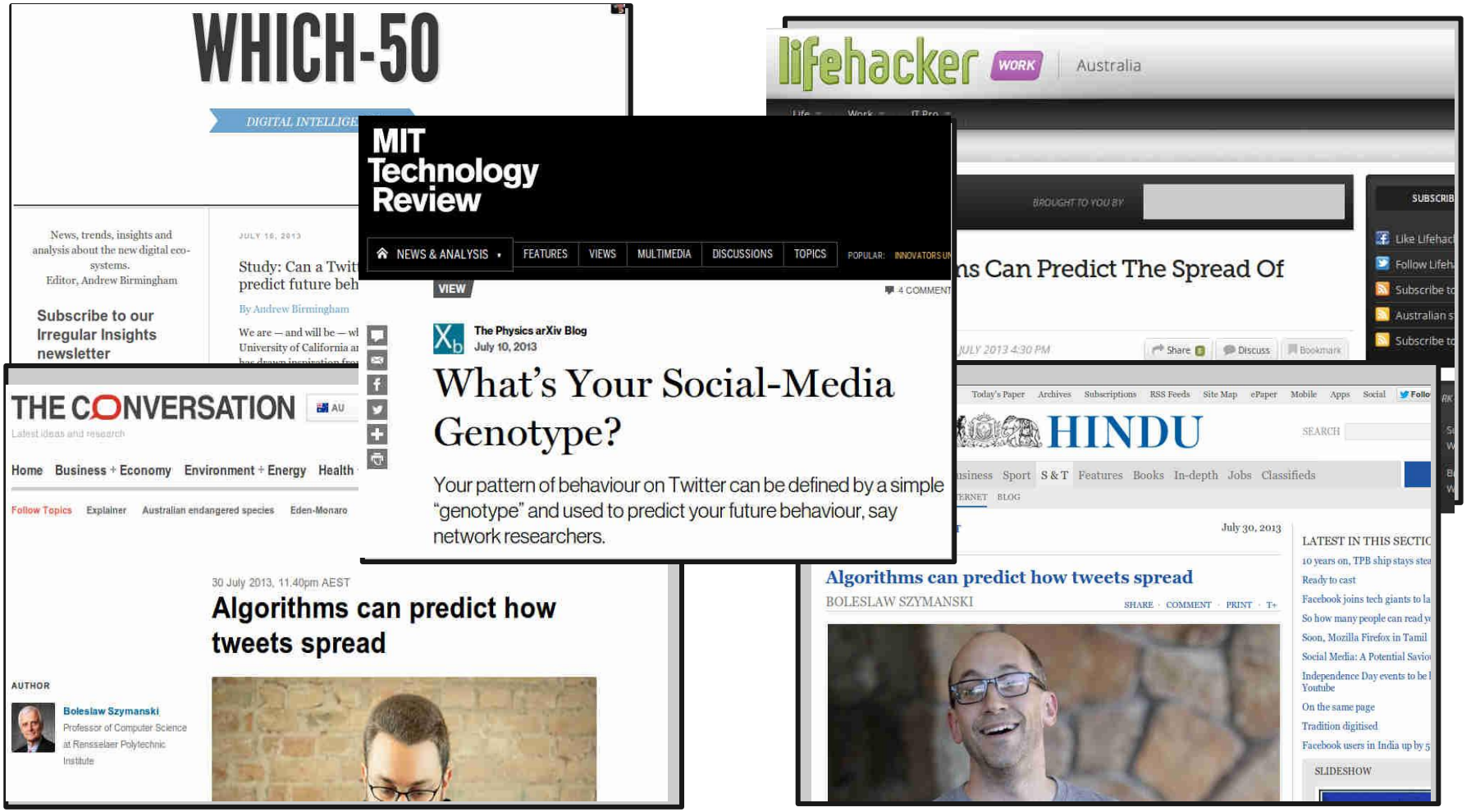


Fig. 6: Comparison of three heuristics for Latency Minimization in the SNAP dataset.

Ongoing and Future Work

- Data sparsity: use urls, sentiment, etc. in addition to HTs.
- Drift of genotypes over long time periods.
- Apply evolution opinion dynamics models to genotypes.
- Estimate exposure rates and size of informed populations using non-linear Kalman Filters.

Media Coverage



Fin

Invariant behavior

On an individual basis, users tend to be consistent in how they respond to a topic.

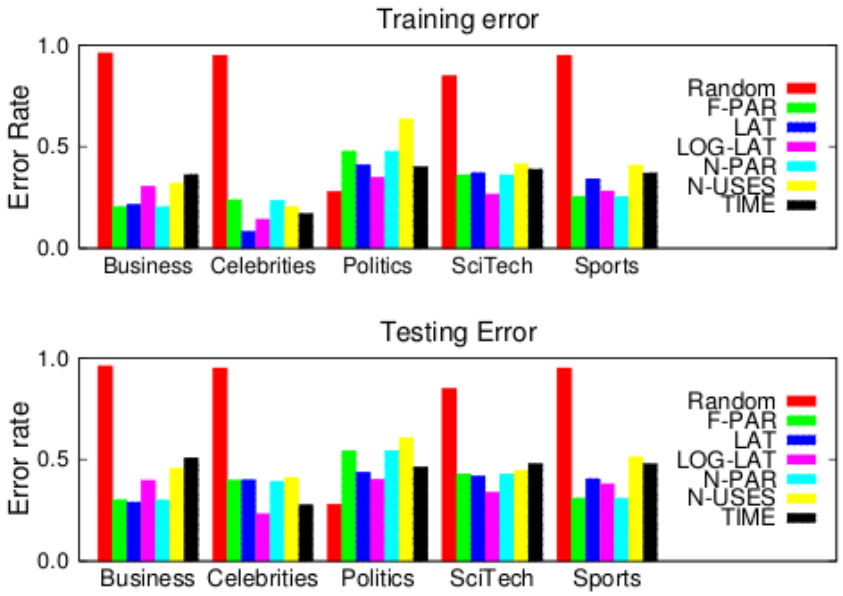
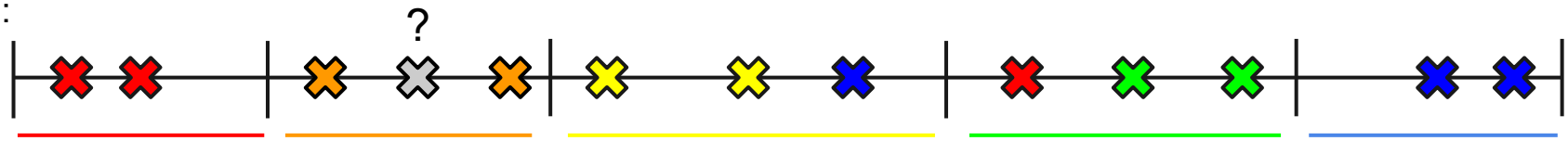


Fig. 1: Training and testing accuracy of leave-one-out Linear Discriminant (LD) classification. MALLET framework text classifier provided HT topic ground truth.

For each user:

Random	Randomly pick topic for HT. Proportional to prior distribution.
F-PAR	Fraction of parents who used HT.
LAT	Inverse of number of posts b/n first HT use of parent and user.
LOG-LAT	Log-normalized version of LAT.
N-USES	Number of HT uses.
TIME	Amount of time b/n first HT use of parent and user.

LD:



Application: Influence Prediction

Goal: *Determine which followees are likely to influence a given user to adopt a hashtag of a certain topic, and, analogously which followers are likely to adopt a hashtag.*

Strategy:

1. Look at the local network structure of a novel hashtag user.
2. Rank that user's set of Topical Influence Network followers based on their propensity to adopt the novel hashtag.
3. Propensities are determined by follower's local network structure, and activity features (genotype).
4. Compare predictions to actual hashtag usage.

Application: Network Latency Minimization

Goal: *Determine which nodes in the topic influence backbone should be targeted for latency reductions, so as to reduce the average minimum latency over the network.*

Strategy:

1. Compute TIME measure of genotype for each node.
2. Discover the backbone for a desired topic.
3. Compute minimum path latencies (sum of TIMEs) between each node.
4. Solve k-LatMin problem (NP-hard) for desired number of target nodes. Set target node latency to zero.