

Detecting polarized structures in social media

Bruno Ordozgoiti

Integrity workshop, WSDM, March 12, 2021, Jerusalem



The dawn of Cyberbalkanization¹.

“It is some time in the future. Technology has greatly increased people’s ability to “filter” what they want to read, see, and hear. [...] With the aid of a television or computer screen, and the Internet, you are able to design your own newspapers and magazines. Having dispensed with broadcasters, you can choose your own video programming, with movies, game shows, sports, shopping, and news of your choice. You mix and match.”

Excerpt from Chapter 1: *“the daily me”*.



¹Sunstein, Cass R. (2001). Republic.com. Princeton: Princeton University Press.

Polarized structures?

- ▶ Filter bubbles,
- ▶ Echo chambers,
- ▶ Extremism,
- ▶ Radicalization...

We will discuss algorithmic ideas to *detect* and *mitigate* these phenomena.

THE NEW YORK TIMES BESTSELLER

THE FILTER BUBBLE

What the Internet is
Hiding from You



'Astonishing'
Andrew Marr

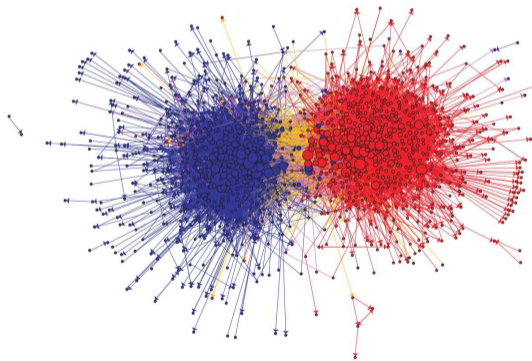
'Explosive'
Chris Anderson

ELI PARISER 

Detection

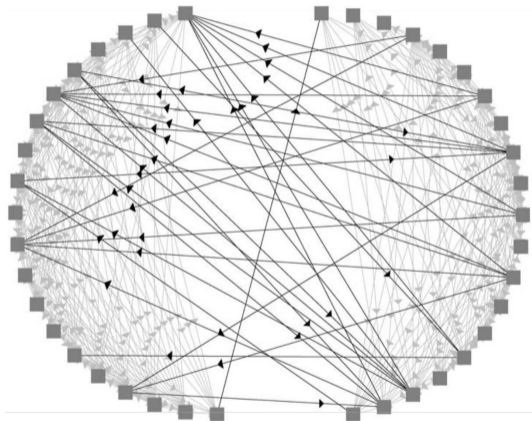
An early example of filter bubbles: link networks between political blogs prior to the 2004 US election (Adamic and Glance, 2005).

"In fact, 91% of the links originating within either the conservative or liberal communities stay within that community²"



²Adamic, Lada A., and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog." Proceedings of the 3rd international workshop on Link discovery. 2005.

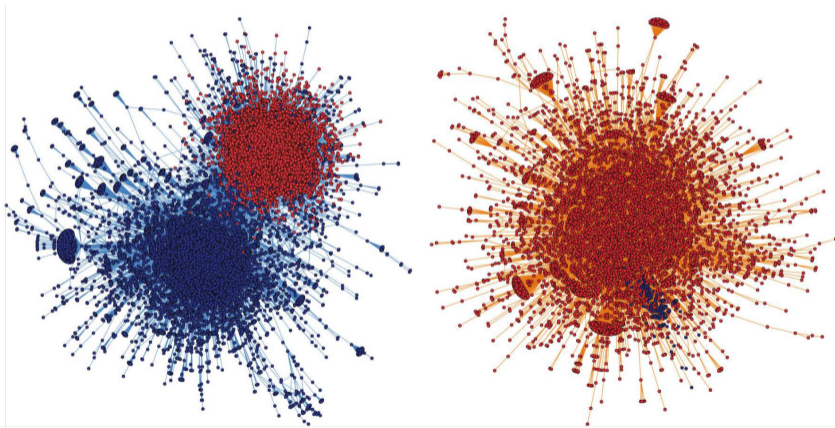
“..there is also some amount of linking to opposing points-of-view. [...] numerous links substantively engage others’ arguments³” (Hargittai et al., 2008)



³Hargittai, Eszter, Jason Gallo, and Matthew Kane. "Cross-ideological discussions among conservative and liberal bloggers." *Public Choice* 134.1-2 (2008): 67-86.

How the network is built is crucial (Conover et al., 2011a):

*“Community structure is evident in the **retweet network**, but less so in the **mention network**.⁴”*



⁴Conover, Michael, et al. "Political polarization on twitter." AAAI WSM 2011.

Direct application of known methods might fail: “...not clear how much modularity is “enough” to state that a social network is polarized [...] on a non-polarized network, cross-group interactions should be at least as frequent as interactions with internal nodes on the community.”⁷”

(Guerra et al., 2013)

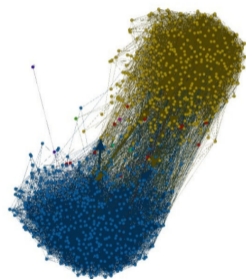
Definitions

Boundary: nodes interacting with the other community.

Polarization: a measure of how much boundary interactions stay within the community.



Facebook friends.
Graduates-undergraduates.
Modularity: 0.24. Polarization: -0.24.



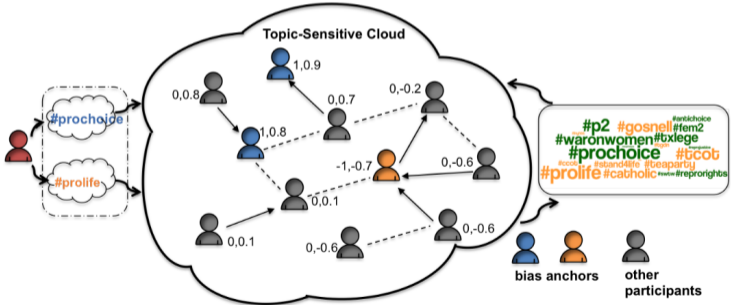
Political blogs.
Liberals-conservatives.
Modularity: 0.48. Polarization: 0.18.

⁷Guerra, Pedro, et al. "A measure of polarization on social media networks based on community boundaries." AAAI WSM 2013.

Biaswatch: a pipeline to detect opinion bias (Lu et al., 2015)

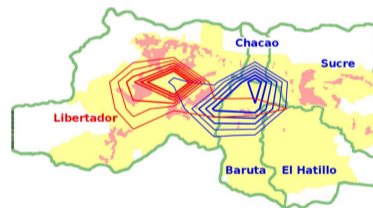
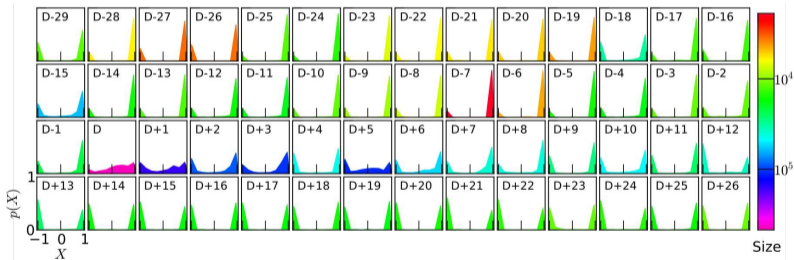
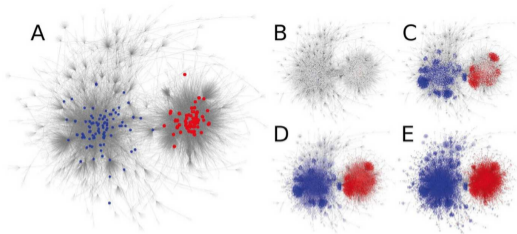
- 1. Find strongly biased users,
- 2. Propagate bias,
- 3. Optimize.

“ Overall, we see a significant improvement of 20.0% in accuracy and 28.6% in AUC on average over the next-best method⁸. ”



⁸Lu, Haokai, James Caverlee, and Wei Niu. "Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media." CIKM 2015.

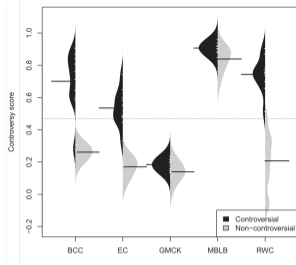
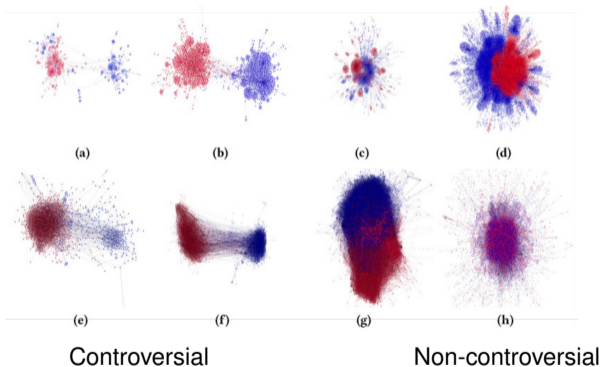
Polarization around death of Hugo Chávez
 (Morales et al., 2015).
 Choice of influential users + DeGroot opinion
 formation model in retweet network.



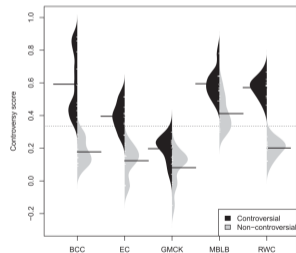
“Which topics spark the most heated debates on social media?” (Garimella et al., 2018)

Random walk controversy

$$RWC = P_{XX}P_{YY} - P_{XY}P_{YX}.$$



Retweet networks - controversy scores.

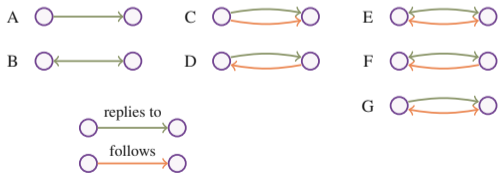


Follow networks - controversy scores.

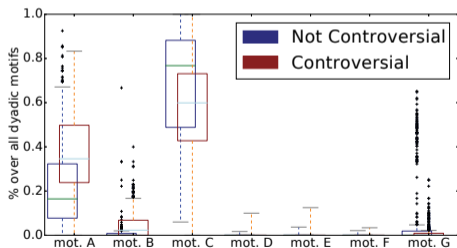
Motif-based controversy detection (Coletto et al., 2017).

Classifier with features from follow + reply graphs

Filtering	Accuracy	Precision	Recall	F-measure
<i>Baseline</i>				
>2 users	0.76	0.79	0.81	0.80
>3 users	0.77	0.80	0.82	0.81
>10 users	0.78	0.81	0.83	0.82
<i>Baseline + dyadic motifs</i>				
>2 users	0.82	0.84	0.86	0.85
>3 users	0.83	0.85	0.86	0.85
>10 users	0.84	0.86	0.88	0.87
<i>Baseline + dyadic and triadic motifs</i>				
>2 users	0.83	0.85	0.86	0.85
>3 users	0.84	0.86	0.85	0.86
>10 users	0.85	0.87	0.88	0.87
<i>Dyadic motifs only</i>				
>2 users	0.75	0.77	0.82	0.80
>3 users	0.75	0.77	0.82	0.80
>10 users	0.77	0.79	0.84	0.82



(a)



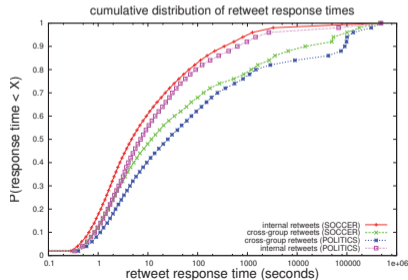
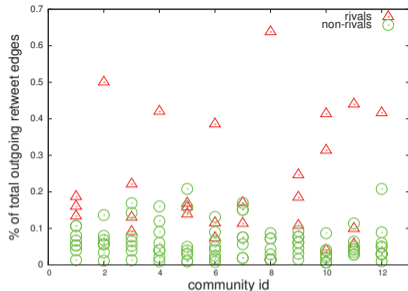
(b)

Check your assumptions!

"...we empirically demonstrate that groups holding antagonistic views can actually retweet each other more often than they retweet other groups⁹."
 (Guerra et al., 2017)

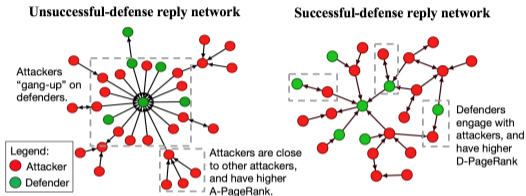
Table 2: Local rivalries in Brazilian Soccer.

Brazilian state	local rivalries
M. Gerais	Cruzeiro, Atlético
S. Paulo	SPFC, Santos, Corint., Palmeiras
R. G. do Sul	Grêmio, Internacional
R. de Janeiro	Flamengo, Flumin., Vasco, Botafogo

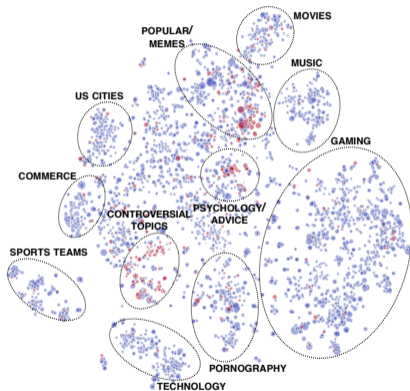
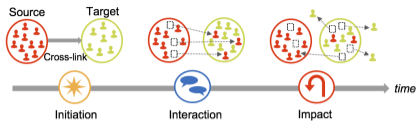


⁹Guerra, Pedro, et al. "Antagonism also flows through retweets: The impact of out-of-context quotes in opinion polarization analysis." Proceedings of the International AAAI WSM 2017.

A look at conflict between Reddit communities. (Kumar et al., 2018)

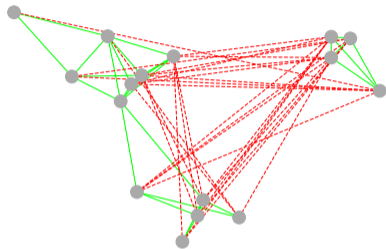


- ▶ "74% of negative mobilizations are initiated by 1% of source communities¹⁰."
- ▶ "...a more fierce defense may be a more effective mitigation strategy, compared to ignoring or isolating the attacking users."



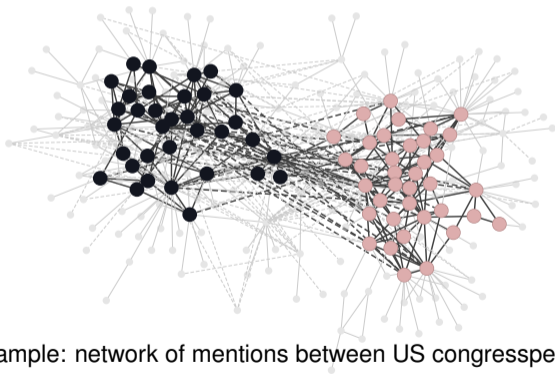
¹⁰Kumar, Srijan, et al. "Community interaction and conflict on the web." WWW 2018.

Methods based on signed networks
(+ and - edges).

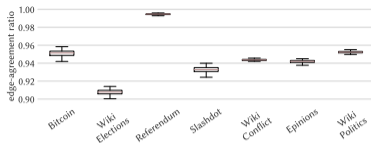


Highland tribes data set.

Community detection + partitioning based on the “first” eigenvector of the signed adjacency matrix (limited to 2 groups).
(Bonchi et al., 2019)



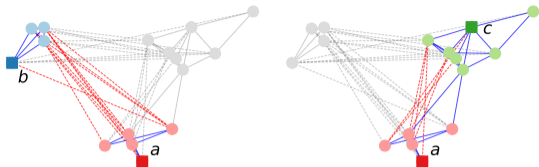
Example: network of mentions between US congresspeople



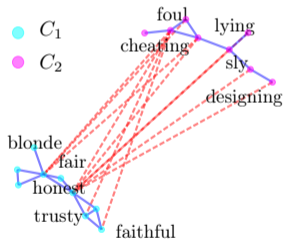
Detected communities are highly polarized

Later extended to k groups by (Tzeng et al., 2020).

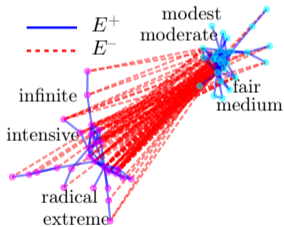
Finding polarized groups with queries (Xiao et al., 2020)



(a) query $S_1 = \{a\}$ and $S_2 = \{b\}$ (b) query $S_1 = \{a\}$ and $S_2 = \{c\}$



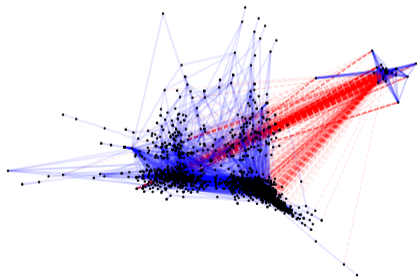
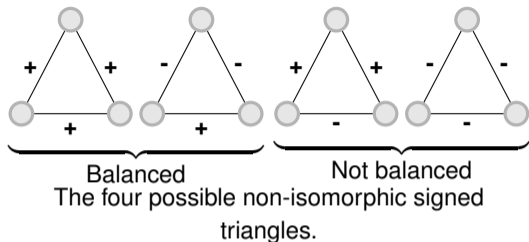
(a) *fair as without cheating*
e.g., a *fair* game



(b) *fair as not excessive*
e.g., *fair* amount of time

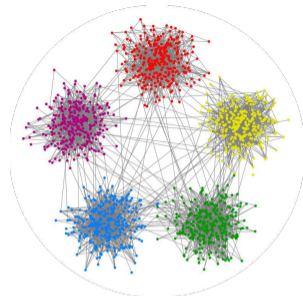
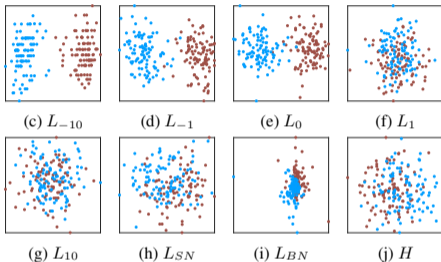
Applications beyond polarization:
synonym/antonym network.

Finding balanced subgraphs
(Ordozgoiti et al., 2020).



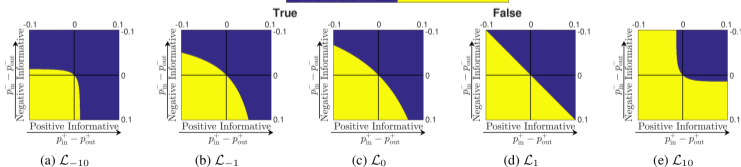
A perfectly polarized subgraph in the Bitcoin trust network.

Recovering signed stochastic block model using power mean Laplacians (Mercado et al., 2019).



(Abbe, 2017)

Recovery of Clusters in Expectation



Mitigation

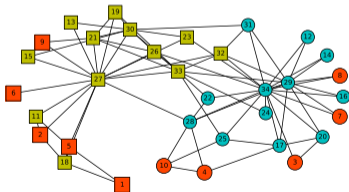
Problem definition (Matakos et al., 2017)

Given a Friedkin-Johnsen opinion network, find the best k moderators.

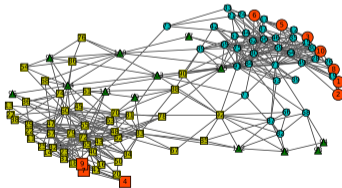
Opinion $\in [-1, 1]$.

A node's opinion is a function of its neighbours'.

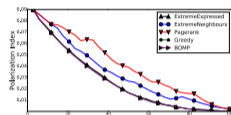
Moderator's opinion = 0.



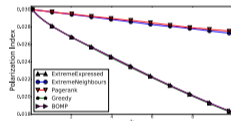
(a)



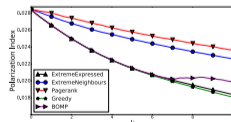
(b)



(a)



(c)

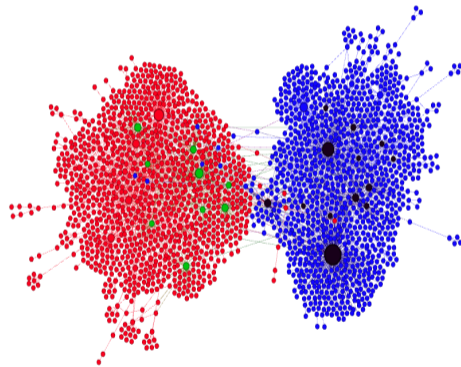
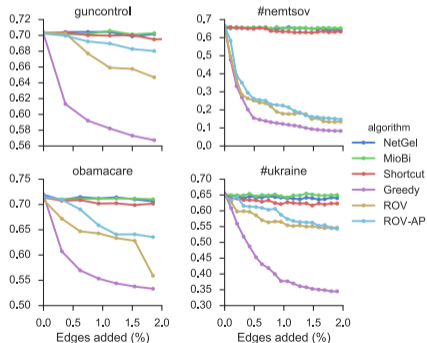


(e)

Problem definition (Garimella et al., 2017a)

Given a network, add k edges to reduce Random Walk Controversy as much as possible.

- ▶ Efficient algorithm¹¹, faster than $\mathcal{O}(n^2)$.
- ▶ Possibility of rejection taken into account.



Nodes chosen in the #russia_march retweet graph.

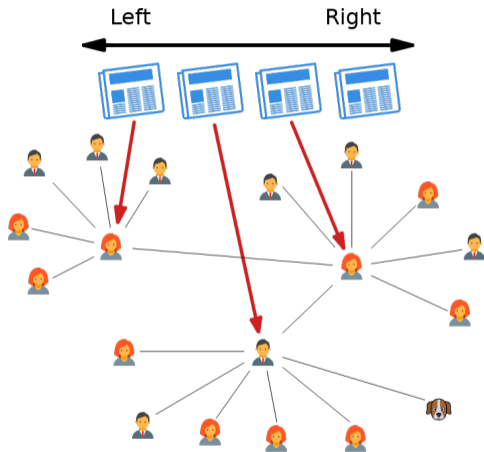
Problem definition (Matakos et al., 2020a)

Given a network of users and a set of items, all with ideological leanings, recommend k items to a set of users so as to maximize exposure diversity across the network.

Assumptions:

- ▶ Propagation is an independent cascade.
- ▶ Users unlikely to spread ideologically-far items.
- ▶ Monotone submodular obj. s.t. matroid constraint.
- ▶ Efficient alg. with modified reverse-reachable sets.

Related work: (Matakos et al., 2020b; Garimella et al., 2017b; Becker et al., 2020)



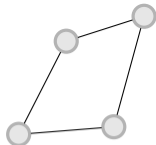
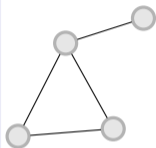
Problem definition (Musco et al., 2018)

Given n agents and an opinion dynamics model, what is the network structure with given weight that minimizes polarization and disagreement simultaneously?

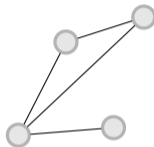
$$\min_L z^T z + z^T L z$$

$$\text{s.t. } L \in \mathcal{L}$$

$$\text{Tr}(L) = 2m$$



?



"Should a recommender system prefer a link suggestion between two users with similar mindsets to keep disagreement low, or two users with different opinions in order to expose each to the others viewpoint of the world?"

Result

There is always a graph with $\mathcal{O}(n/\epsilon^2)$ edges that achieves a $(1 + \epsilon)$ -approximation.

Conflict **risk** minimization in a Friedkin-Johnsen model (each user i has an internal s_i and an expressed opinion z_i) (Chen et al., 2018).

Measures of conflict :

- ▶ Internal conflict: $\sum_i (z_i - s_i)^2$.
- ▶ External conflict: $\sum_{(i,j) \in E} w_{ij} (z_i - z_j)^2$.
- ▶ Controversy: $\sum_i z_i^2$.
- ▶ Resistance: $r = z^T s$.

Conservation law of conflict

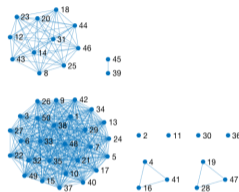
$$IC + 2 \times EC + C = s^T s = \sum_i s_i^2.$$

Proposal: minimize **expected** and **worst-case** conflict risk over s , w.r.t. edge editions.

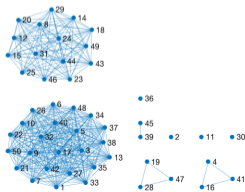
(a) ACR = 8.4116



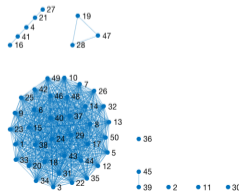
(b) ACR = 2.6315



(c) ACR = 2.6710



(d) ACR = 2.3308



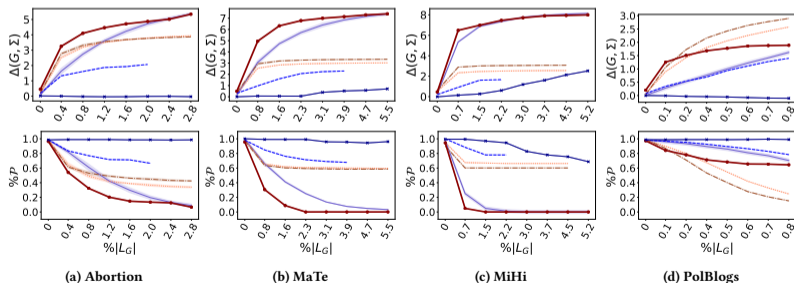
Definitions

Consider a graph with vertices of 2 colours.

- ▶ Bubble radius of v : hitting time of a vertex of a different colour from v .
- ▶ Parochials: nodes with high BR.
- ▶ Cosmopolitans: nodes with low BR.

Problem: add k edges to maximally reduce n. of parochials.

Results: monotone submodular objective. Approx. guarantees.



— PB ···· 10-RCN - - - 10-WRCN — RePBubLik - - - ROV — node2vec

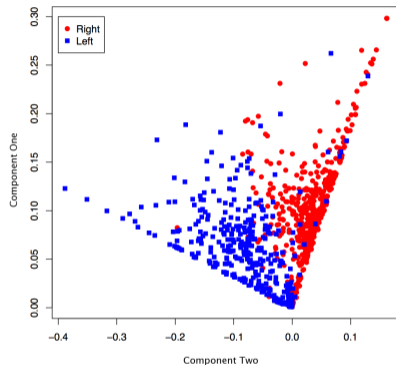
Annotation

Predicting political alignment on Twitter (Conover et al., 2011b)

SVM trained on different feature vectors:

- ▶ Tweet text
- ▶ Hashtags
- ▶ Network structure

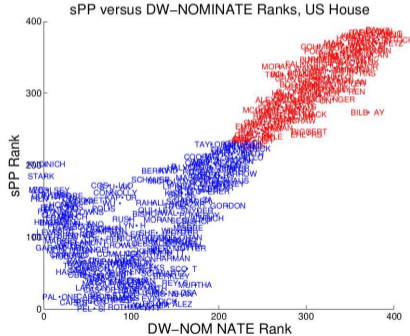
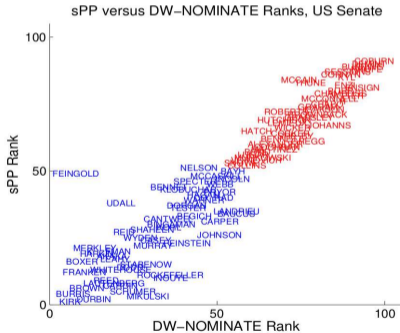
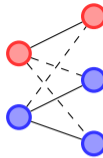
Features	Conf. matrix	Accuracy	Section
Full-Text	$\begin{bmatrix} 266 & 107 \\ 75 & 431 \end{bmatrix}$	79.2%	§ IV-A1
Hashtags	$\begin{bmatrix} 331 & 42 \\ 41 & 465 \end{bmatrix}$	90.8%	§ IV-A2
Clusters	$\begin{bmatrix} 367 & 6 \\ 38 & 468 \end{bmatrix}$	94.9%	§ IV-B
Clusters + Tags	$\begin{bmatrix} 366 & 7 \\ 38 & 468 \end{bmatrix}$	94.9%	§ IV-B



First two LSA factors of the hashtag-user matrix.

Node classification and ranking based on signed bipartite network¹².

Markov random field MLE with loopy belief propagation.
(Akoglu, 2014)

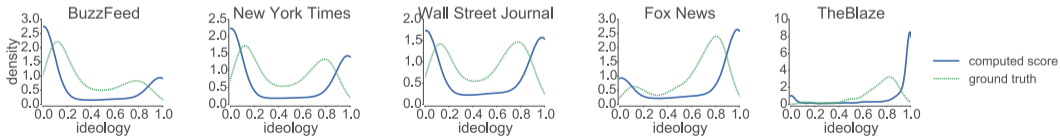
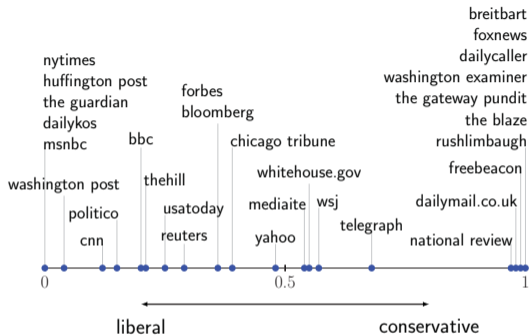


¹²Akoglu, Leman. "Quantifying political polarity based on bipartite opinion networks." Proceedings of the International AAAI Conference on Web and Social Media, 2014.

User and source ideology with joint NMF (Lahoti et al., 2018).

Setting

- ▶ User-user matrix $A \approx UH_uU^T$.
- ▶ User-source matrix $C \approx UH_sV^T$.



“...reported accuracies have been systemically over-optimistic due to the way in which validation datasets have been collected¹³. (Cohen and Ruths, 2013)”

Previous works focus on highly-political accounts (politicians or self-declared users). This work:

- ▶ US politicians,
- ▶ users with self-reported political orientation,
- ▶ modest users who do not declare their political views, but make sufficient mention of politics.

Result

- ▶ Performance drops from $\geq 90\%$ to 65%.
- ▶ *“We found that classifiers based on politicians, while achieving 91% labeling accuracy on other politicians, only achieved 11% accuracy on politically modest users.”*

¹³Cohen, Raviv, and Derek Ruths. “Classifying political orientation on Twitter: It’s not easy!” AAAI WSM.

Finding early signs of Trump support on Reddit (Massachs et al., 2020).

“We find that homophily-based and social feedback-based features are the most predictive signals.”

	Precision	Recall	F1	AUC
Participation	0.25	0.56	0.34	0.68
Score	0.24	0.60	0.33	0.67
Interaction	0.18	0.52	0.26	0.55
Part. + Score	0.27	0.56	0.35	0.70

Trump supporters		Trump non-supporters	
Subreddit	β	Subreddit	β
r/Conservative	0.3815	r/raspberry_pi	-0.2847
r/Libertarian	0.3740	r/TrueAtheism	-0.2577
r/conspiracy	0.3733	r/AskCulinary	-0.2355
r/4chan	0.3341	r/comics	-0.2249
r/circlejerk	0.3107	r/rpg	-0.2186
r/NoFap	0.2918	r/ireland	-0.2034
r/Entrepreneur	0.2539	r/Fantasy	-0.1983
r/ImGoingToHellForThis	0.2510	r/explainlikeimfive	-0.1944
r/trees	0.2482	r/environment	-0.1892
r/MensRights	0.2482	r/doctorwho	-0.1878
r/guns	0.2293	r/polyamory	-0.1806
r/blackops2	0.2110	r/scifi	-0.1777
r/runescape	0.2031	r/books	-0.1772
r/Anarcho_Capitalism	0.1937	r/askscience	-0.1738
r/Catholicism	0.1931	r/london	-0.1691
r/leagueoflegends	0.1920	r/britishproblems	-0.1687
r/nfl	0.1843	r/Homebrewing	-0.1632
r/starcraft	0.1714	r/programming	-0.1521
r/CCW	0.1638	r/gadgets	-0.1501
r/breakingbad	0.1631	r/AndroidQuestions	-0.1463
r/investing	0.1624	r/listentothis	-0.1462
r/AdviceAnimals	0.1589	r/hiphopheads	-0.1397
r/DeadBedrooms	0.1577	r/boardgames	-0.1336
r/Firearms	0.1551	r/asoiaf	-0.1292
r/Advice	0.1537	r/whatisthisthing	-0.1244
r/seduction	0.1518	r/lgbt	-0.1187
r/Christianity	0.1455	r/cringepics	-0.1175
r/golf	0.1453	r/ukpolitics	-0.1136
r/mylittlepony	0.1437	r/Python	-0.1089
r/POLITIC	0.1423	r/baseball	-0.1080

A word on signed networks:

- ▶ There is a shortage of signed networks! E.g. check:
 - ▶ <https://snap.stanford.edu/data/> 9 out of about 120 are signed.
 - ▶ <http://konect.cc/> 8 out of 1,326 are signed.
- ▶ Annotating user interactions with signs is challenging. Ideas are welcome!



To conclude, a personal note: we won't solve this with algorithms alone.

Thanks for listening, and thanks to all researchers working in the field, especially present and past members of the group led by Aris Gionis at Aalto/KTH.

bruno.ordozgoiti [at] aalto.fi



References I

- Abbe, E. (2017). Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531.
- Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43.
- Akoglu, L. (2014). Quantifying political polarity based on bipartite opinion networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.
- Becker, R., Coro, F., D’Angelo, G., and Gilbert, H. (2020). Balancing spreads of influence in a social network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3–10.

References II

- Bonchi, F., Galimberti, E., Gionis, A., Ordozgoiti, B., and Ruffo, G. (2019). Discovering polarized communities in signed networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 961–970.
- Chen, X., Lijffijt, J., and De Bie, T. (2018). Quantifying and minimizing risk of conflict in social networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1197–1205.
- Cohen, R. and Ruths, D. (2013). Classifying political orientation on twitter: It's not easy! In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7.
- Coletto, M., Garimella, K., Gionis, A., and Lucchese, C. (2017). A motif-based approach for identifying controversy. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

References III

- Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. (2011a). Political polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.
- Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. (2011b). Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 192–199. IEEE.
- Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2017a). Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 81–90.
- Garimella, K., Gionis, A., Parotsidis, N., and Tatti, N. (2017b). Balancing information exposure in social networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4666–4674.

References IV

- Garimella, K., Morales, G. D. F., Gionis, A., and Mathioudakis, M. (2018). Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27.
- Gionis, A., Terzi, E., and Tsaparas, P. (2013). Opinion maximization in social networks. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 387–395. SIAM.
- Guerra, P., Meira Jr, W., Cardie, C., and Kleinberg, R. (2013). A measure of polarization on social media networks based on community boundaries. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7.
- Guerra, P., Nalon, R., Assunção, R., and Meira Jr, W. (2017). Antagonism also flows through retweets: The impact of out-of-context quotes in opinion polarization analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

References V

- Haddadan, S., Menghini, C., Riondato, M., and Upfal, E. (2021). Republik: Reducing the polarized bubble radius with link insertions. *arXiv preprint arXiv:2101.04751*.
- Hargittai, E., Gallo, J., and Kane, M. (2008). Cross-ideological discussions among conservative and liberal bloggers. *Public Choice*, 134(1-2):67–86.
- Kumar, S., Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2018). Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference*, pages 933–943.
- Lahoti, P., Garimella, K., and Gionis, A. (2018). Joint non-negative matrix factorization for learning ideological leaning on twitter. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 351–359.

References VI

- Lu, H., Caverlee, J., and Niu, W. (2015). Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 213–222.
- Massachs, J., Monti, C., Morales, G. D. F., and Bonchi, F. (2020). Roots of trumpism: Homophily and social feedback in donald trump support on reddit. In *12th ACM Conference on Web Science*, pages 49–58.
- Matakos, A., Aslay, C., Galbrun, E., and Gionis, A. (2020a). Maximizing the diversity of exposure in a social network. *IEEE Transactions on Knowledge and Data Engineering*.
- Matakos, A., Terzi, E., and Tsaparas, P. (2017). Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery*, 31(5):1480–1505.

References VII

- Matakos, A., Tu, S., and Gionis, A. (2020b). Tell me something my friends do not know: Diversity maximization in social networks. *Knowledge and Information Systems*, 62(9):3697–3726.
- Mercado, P., Tudisco, F., and Hein, M. (2019). Spectral clustering of signed graphs via matrix power means. In *International Conference on Machine Learning*, pages 4526–4536. PMLR.
- Morales, A. J., Borondo, J., Losada, J. C., and Benito, R. M. (2015). Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3):033114.
- Musco, C., Musco, C., and Tsourakakis, C. E. (2018). Minimizing polarization and disagreement in social networks. In *Proceedings of the 2018 World Wide Web Conference*, pages 369–378.

References VIII

- Ordozgoiti, B., Matakos, A., and Gionis, A. (2020). Finding large balanced subgraphs in signed networks. In *Proceedings of The Web Conference 2020*, pages 1378–1388.
- Tzeng, R.-C., Ordozgoiti, B., and Gionis, A. (2020). Discovering conflicting groups in signed networks. *Advances in Neural Information Processing Systems*, 33.
- Xiao, H., Ordozgoiti, B., and Gionis, A. (2020). Searching for polarization in signed graphs: a local spectral approach. In *Proceedings of The Web Conference 2020*, pages 362–372.