The Low Hanging Fruit of the Twitter Following Graph

Alex Hayes

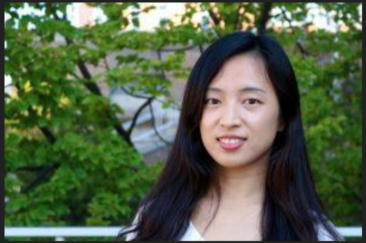
August 11, 2021

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

@alexpghayes

2021 JSM
Statistics,
the Storie
Virtual C
August 8–12

This is joint work



Yini Zhang

Assistant Professor,
U Buffalo

Nathan Kolbow

Incoming PhD Student,
UW-Madison

Fan Chen

Data Scientist,
Google

Karl Rohe

Professor,
UW-Madison

Observational research on Twitter

Large body of work using tweets

Comparatively little work using the following graph

This talk:

1. Why empirical work using the following graph is hard
2. Tools to make it easier
3. The value of the following graph

Why applied work using the following graph is hard

Twitter's public API rate limited to 5,000 edges/minute

The following graph is huge (~350 million monthly active users, ~200 friends/user)

Implications:

1. Can’t waste any API requests
2. Need to cache data for robustness in long running API requests
3. Cached data needs to support graph queries for adaptive sampling

Till now, very little infrastructure to support this type of data collection

neocache is a tool to cache the following graph

User perspective: Drop in replacements for *rtweet* functionality

rtweet::lookup_users() → neocache::nc_lookup_users()

rtweet::get_friends() → neocache::nc_get_friends()

Developer perspective:

- Data cached in Neo4J database running inside Docker container
- O(1) neighborhood lookups
- Complex caching logic due to many forms of partial information availability

https://github.com/alexpghayes/neocache

Sampling the following graph

Our sampling strategy:

- Known seed nodes of interest
- Want the local network around these

Can we snowball sample?

Sampling the following graph

Our sampling strategy:

- Known seed nodes of interest
- Want the local network around these

Can we snowball sample? No.

Not enough data

Exceeds API limits

1-hop neighborhood: ~1000 nodes
2-hop neighborhood: ~1,000,000 nodes
3-hop neighborhood: All of Twitter

Personalized PageRank

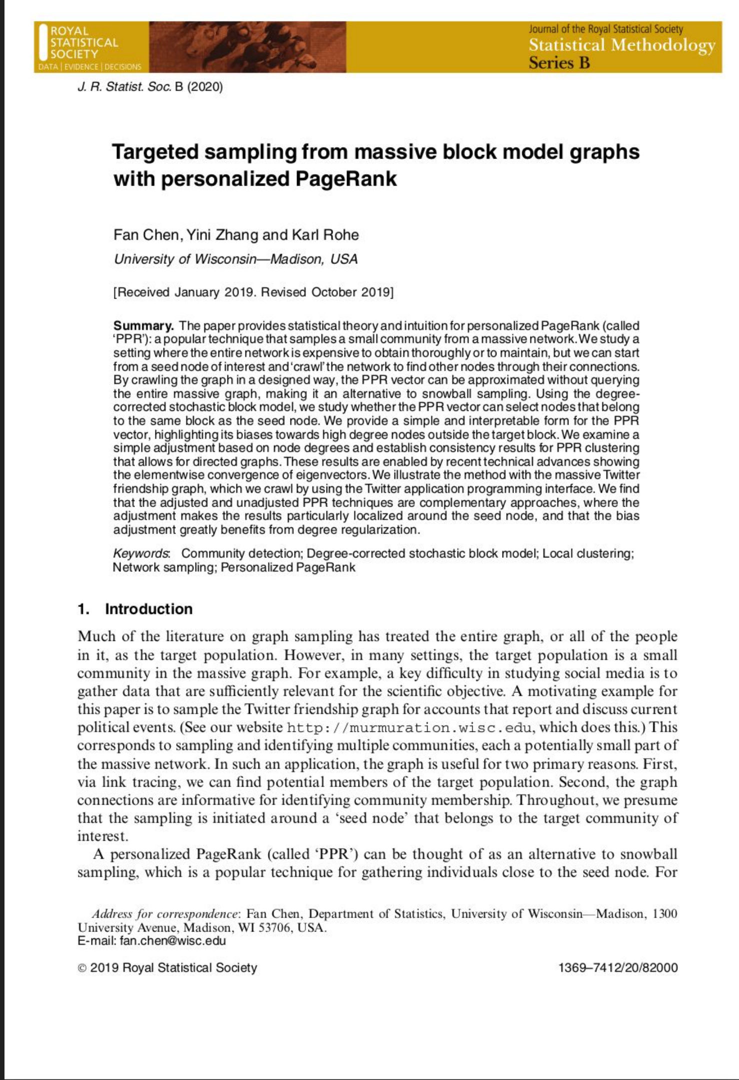
Sample nodes with high Personalized PageRank w.r.t. nodes [1-2]

Compute an ε-approximation

ε determines how much data we need

[1] Andersen, Reid, Fan Chung, and Kevin Lang. "Local Graph Partitioning Using PageRank Vectors." In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, 475–86. Berkeley, CA, USA: IEEE, 2006. https://doi.org/10.1109/FOCS.2006.44.

[2] Chen, Fan, Yini Zhang, and Karl Rohe. "Targeted Sampling from Massive Blockmodel Graphs with Personalized PageRank." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82, no. 1 (February 2020): 99–126. https://doi.org/10.1111/rssb.12349.



ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)

Targeted sampling from massive block model graphs with personalized PageRank

Fan Chen, Yini Zhang and Karl Rohe

University of Wisconsin—Madison, USA

[Received January 2019. Revised October 2019]

Summary. The paper provides statistical theory and intuition for personalized PageRank (called 'PPR'): a popular technique that samples a small community from a massive network. We study a setting where the entire network is expensive to obtain thoroughly or to maintain, but we can start from a seed node of interest and 'crawl' the network to find other nodes through their connections. By crawling the graph in a designed way, the PPR vector can be approximated without querying the entire massive graph, making it an alternative to snowball sampling. Using the degree-corrected stochastic block model, we study whether the PPR vector can select nodes that belong to the same block as the seed node. We provide a simple and interpretable form for the PPR vector, highlighting its biases towards high degree nodes outside the target block. We examine a simple adjustment based on node degrees and establish consistency results for PPR clustering that allows for directed graphs. These results are enabled by recent technical advances showing the elementwise convergence of eigenvectors. We illustrate the method with the massive Twitter friendship graph, which we crawl by using the Twitter application programming interface. We find that the adjusted and unadjusted PPR techniques are complementary approaches, where the adjustment makes the results particularly localized around the seed node, and that the bias adjustment greatly benefits from degree regularization.

Keywords: Community detection; Degree-corrected stochastic block model; Local clustering; Network sampling; Personalized PageRank

1. Introduction

Much of the literature on graph sampling has treated the entire graph, or all of the people in it, as the target population. However, in many settings, the target population is a small community in the massive graph. For example, a key difficulty in studying social media is to gather data that are sufficiently relevant for the scientific objective. A motivating example for this paper is to sample the Twitter friendship graph for accounts that report and discuss current political events. (See our website http://murmuration.wisc.edu, which does this.) This corresponds to sampling and identifying multiple communities, each a potentially small part of the massive network. In such an application, the graph is useful for two primary reasons. First, via link tracing, we can find potential members of the target population. Second, the graph connections are informative for identifying community membership. Throughout, we presume that the sampling is initiated around a 'seed node' that belongs to the target community of interest.

A personalized PageRank (called 'PPR') can be thought of as an alternative to snowball sampling, which is a popular technique for gathering individuals close to the seed node. For

Address for correspondence: Fan Chen, Department of Statistics, University of Wisconsin—Madison, 1300 University Avenue, Madison, WI 53706, USA.
E-mail: fan.chen@wisc.edu

1369–7412/20/82000

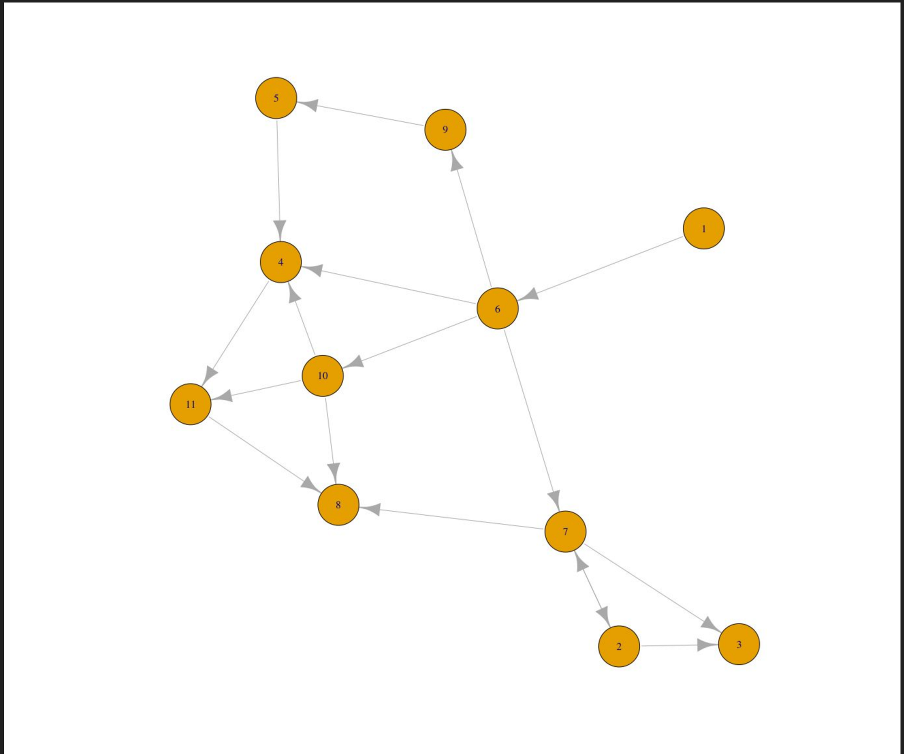
Personalized PageRank

Start at seed node

1. Visit seed with probability α
2. Follow edge with probability 1 - α
    a. Visit seed if there are no edges to follow

Stationary distribution *p* defines PPR

A node has high PPR if there are lots of paths from the seed to that node



Personalized PageRank approximation for directed graphs

Algorithm 3 Approximate PPR Vector (directed)

Require: Directed graph G, preference vector π, teleportation constant α, and tolerance ϵ.

Initialize $p \leftarrow 0$, $r \leftarrow \pi$, $\alpha' \leftarrow \alpha/(2-\alpha)$.
while $\exists u \in V$ such that $r_u \geq \epsilon d_u^{\text{out}}$ **do**
  Sample a vertex u uniformly at random, satisfying $r_u \geq \epsilon d_u^{\text{out}}$.
  $p_u \leftarrow p_u + \alpha' r_u$.
  for $v : (u, v) \in E$ **do**
    $r_v \leftarrow r_v + (1-\alpha')r_u/(2d_u^{\text{out}})$.
  end for
  $r_u \leftarrow (1-\alpha')r_u/2$.
end while

Return: ϵ-approximate PPR vector p.

Chen, Fan, Yini Zhang, and Karl Rohe. “Targeted Sampling from Massive Blockmodel Graphs with Personalized PageRank.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82, no. 1 (February 2020): 99–126. https://doi.org/10.1111/rssb.12349.

Personalized PageRank approximation for directed graphs

Algorithm 3 Approximate PPR Vector (directed)

Require: Directed graph G, preference vector π, teleportation constant α, and tolerance ϵ.

  Initialize $p \leftarrow 0$, $r \leftarrow \pi$, $\alpha' \leftarrow \alpha/(2-\alpha)$.
  while $\exists u \in V$ such that $r_u \geq \epsilon d_u^{\mathrm{out}}$ **do**
    Sample a vertex u uniformly at random, satisfying $r_u \geq \epsilon d_u^{\mathrm{out}}$.
    $p_u \leftarrow p_u + \alpha' r_u$.
    for $v : (u, v) \in E$ **do**
      $r_v \leftarrow r_v + (1-\alpha')r_u/(2d_u^{\mathrm{out}})$.
    end for
    $r_u \leftarrow (1-\alpha')r_u/2$.
  end while

Return: ϵ-approximate PPR vector p.

We don't actually need the full graph G

Chen, Fan, Yini Zhang, and Karl Rohe. "Targeted Sampling from Massive Blockmodel Graphs with Personalized PageRank." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82, no. 1 (February 2020): 99–126. https://doi.org/10.1111/rssb.12349.

Personalized PageRank approximation for directed graphs

Algorithm 3 Approximate PPR Vector (directed)

Require: Directed graph G, preference vector π, teleportation constant α, and tolerance ϵ.

  Initialize $p \leftarrow 0$, $r \leftarrow \pi$, $\alpha' \leftarrow \alpha/(2-\alpha)$.

  while $\exists u \in V$ such that $r_u \geq \epsilon d_u^{\text{out}}$ **do**

    Sample a vertex u uniformly at random, satisfying $r_u \geq \epsilon d_u^{\text{out}}$.

    $p_u \leftarrow p_u + \alpha' r_u$.

    for $v : (u, v) \in E$ **do**

      $r_v \leftarrow r_v + (1 - \alpha')r_u/(2d_u^{\text{out}})$.

    end for

    $r_u \leftarrow (1 - \alpha')r_u/2$.

  end while

Return: ϵ-approximate PPR vector p.

Need node degrees

Chen, Fan, Yini Zhang, and Karl Rohe. "Targeted Sampling from Massive Blockmodel Graphs with Personalized PageRank." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82, no. 1 (February 2020): 99–126. https://doi.org/10.1111/rssb.12349.

Personalized PageRank approximation for directed graphs

Algorithm 3 Approximate PPR Vector (directed)

Require: Directed graph G, preference vector π, teleportation constant α, and tolerance ϵ.

  Initialize $p \leftarrow 0$, $r \leftarrow \pi$, $\alpha' \leftarrow \alpha/(2-\alpha)$.
  while $\exists u \in V$ such that $r_u \geq \epsilon d_u^{\text{out}}$ **do**
    Sample a vertex u uniformly at random, satisfying $r_u \geq \epsilon d_u^{\text{out}}$.
    $p_u \leftarrow p_u + \alpha' r_u$.
    for $v : (u, v) \in E$ **do**
      $r_v \leftarrow r_v + (1-\alpha') r_u / (2 d_u^{\text{out}})$.
    end for
    $r_u \leftarrow (1-\alpha') r_u / 2$.
  end while

Return: ϵ-approximate PPR vector p.

Need ego networks

Chen, Fan, Yini Zhang, and Karl Rohe. “Targeted Sampling from Massive Blockmodel Graphs with Personalized PageRank.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82, no. 1 (February 2020): 99–126. https://doi.org/10.1111/rssb.12349.

aPPR is a tool to approximate Personalized PageRank

User perspective

Computes PPR for arbitrary graph with methods

- *degrees(graph, nodes)*
- *neighborhood(graph, node)*

* In practice Algorithm 3 [last slide] needs an extra method that checks if a node is available via API and ignores that node if it isn't

Developer perspective

Designed for graphs primarily accessible via API:

- Runtime dominated by data transfer over networks
- Carefully implemented to avoid extraneous API requests

** The implementation relies on a mixture of generic function (S3) and more classic encapsulated OOP (R6)

neocache + *aPPR* are well integrated

Approximate Personalized PageRank using aPPR

Tell aPPR to query Twitter API via neocache

Export data from neocache when PPR calculation is done

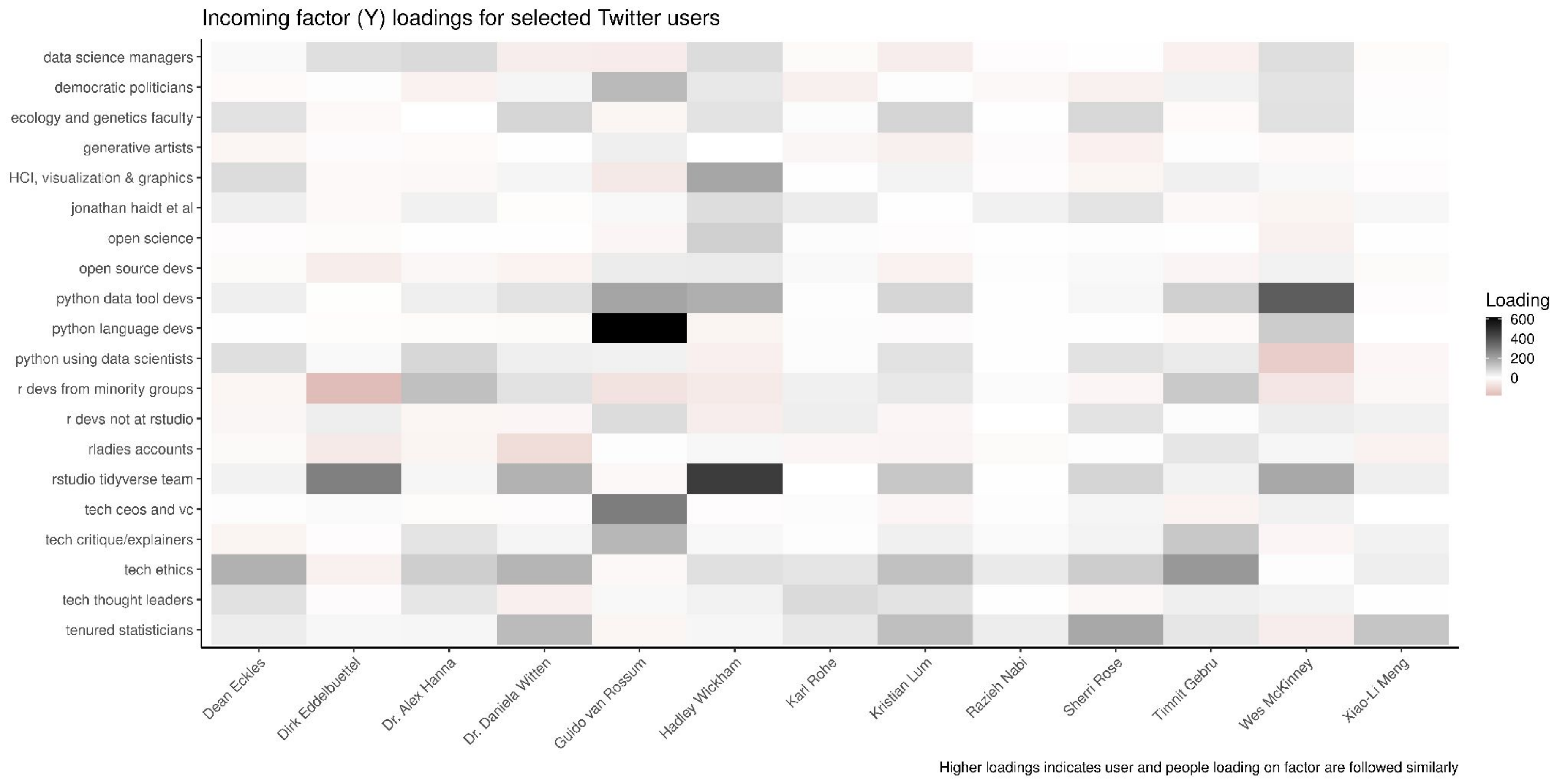
```r
library(aPPR)
library(neocache)

set.seed(26)

# this takes about 33 hours due to API rate limits

tracker <- appr(
  neocache_graph(),
  seed = c("hadleywickham", "gvanrossum"),
  epsilon = 1e-6
)

nc_export_all_follows("aPPR", "path/to/edgelist")
nc_export_all_users("aPPR", "path/to/nodelist")
```

The following graph is a high signal dataset

Method

1. Calculate Personalized PageRanks seeded at @hadleywickham + @gvanrossum
2. Get all outgoing edges from users with high Personalized PageRanks
3. Take a rank 20 SVD of the adjacency matrix $A \approx U D V'$
4. Varimax rotate U and V to obtain $A \approx Z B Y'$

Factor	Name	Keywords
Y01	big hitters in statistics/ASA	statistics, professor, biostatistics, statistical, statistician, amherst, data
Y02	data science managers	data, dc, analytics, scientist, science, nyc, #rstats
Y03	democratic politicians	the, us, author, official, senator, news, host
Y04	ecology and genetics faculty	evolutionary, genetics, evolution, genomics, biologist, population, biology
Y05	generative artists	design, designer, art, artist, graphics, creative, generative
Y06	HCI, visualization & graphics	visualization, data, professor, hci, graphics, design, visual
Y07	jonathan haidt et al	professor, political, visualization, phd, author, prof, evolutionary
Y08	open science	open, science, research, scholarly, publishing, access, #openscience
Y09	open source devs	zealand, science, open, @thecarpentries, @johndcook, aotearoa, nz
Y10	python data tool devs	data, machine, python, learning, ai, science, scientist
Y11	python language devs	python, developer, software, @thepsf, django, core, engineer
Y12	python using data scientists	data, @etsy, scientist, #rstats, sheher, etsy, @nytimes
Y13	r devs from minority groups	data, #rstats, sheher, scientist, r, science, |
Y14	r devs not at rstudio	#rstats, data, r, @rstudio, scientist, rstudio, hehim
Y15	rladies accounts	#rstats, r, #rladies, rladies, diversity, data, gender
Y16	rstudio tidyverse team	#rstats, data, r, @rstudio, science, scientist, statistics
Y17	tech ceos and vc	investor, @dropbox, ceo, cofounder, founder, dropbox, google
Y18	tech critique/explainers	sheher, theythem, hehim, queer, security, i, infosec
Y19	tech ethics	professor, prof, ai, phd, research, assistant, machine
Y20	tech thought leaders	cofounder, ceo, ai, vc, founder, tech, data


Incoming factor (Y) loadings for selected Twitter users
data science managers
democratic politicians
ecology and genetics faculty
generative artists
HCI, visualization & graphics
jonathan haidt et al
open science
open source devs
python data tool devs
python language devs
python using data scientists
r devs from minority groups
r devs not at rstudio
rladies accounts
rstudio tidyverse team
tech ceos and vc
tech critique/explainers
tech ethics
tech thought leaders
tenured statisticians
Dean Eckles
Dirk Eddelbuettel
Dr. Alex Hanna
Dr. Daniela Witten
Guido van Rossum
Hadley Wickham
Karl Rohe
Kristian Lum
Razieh Nabi
Sherri Rose
Timnit Gebru
Wes McKinney
Xiao-Li Meng
Loading
600
400
200
0
Higher loadings indicates user and people loading on factor are followed similarly

Thank you! Questions?

@alexpghayes

@alexpghayes

alex.hayes@wisc.edu

aPPR: https://github.com/alexpghayes/neocache *

neocache: https://github.com/RoheLab/aPPR *

Not convinced about the following graph? Play with the data yourself https://github.com/alexpghayes/JSM2021 **

* Documentation currently lags functionality, tackling this very soon
** Be sure to read the LICENSE section of the included README

Appendix

The following graph is consistently a high signal dataset

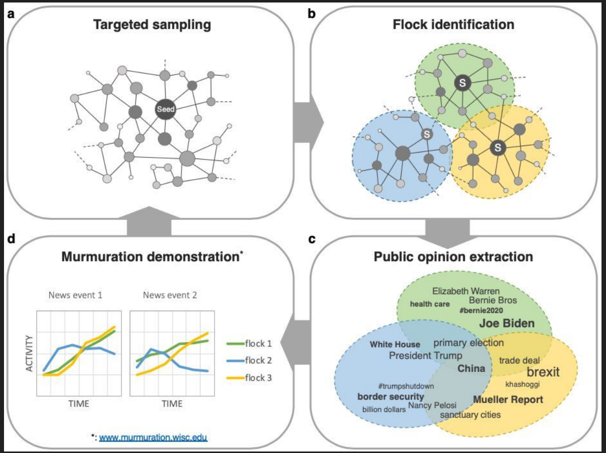
Zhang, Yini, Fan Chen, and Karl Rohe. “Social Media Public Opinion as Flocks in a Murmuration: Conceptualizing and Measuring Opinion Expression on Social Media.” *Journal of Computer-Mediated Communication*, 2021+.

Several other ongoing projects:

- Identifying trustworthy Twitter users
- Finding high quality botnets where bots are sometime run by humans
- Information pathways between users

Table 3: Top 30 handles of PPR with seed node @NBCPolitics and the teleportation constant $\alpha = 0.15$ in December 2018.

	Name	Followers	Description
1	Melania Trump	11242283	This account is run by the Office of First Lady Melania Trump...
2	The White House	17625630	Welcome to @WhiteHouse! Follow for the latest from President...
3	Chuck Todd	2032038	Moderator of @meetthepress and @nbcnews political director; ...
4	NBC News	6280551	The leading source of global news and info for more than 75 ...
5	NBC Nightly News	962290	Breaking news, in-depth reporting, context on news from ...
6	Andrea Mitchell	1737764	NBC News Chief Foreign Affairs Correspondent/anchor, Andrea ...
7	Savannah Guthrie	881669	Mom to Vale & Charley, TODAY Co-Anchor, Georgetown Law. ...
8	Joe Scarborough	2521215	With Malice Toward None
9	MSNBC	2261911	The place for in-depth analysis, political commentary and ...
10	Rachel Maddow MSNBC	9498076	I see political people...



Factor	Name	Top Accounts
Y01	big hitters in statistics/ASA	Elizabeth Stuart, ASA, Michael Love, Dr. Leslie McClure, Sherri Rose, francesca dominici, Emma Benn
Y02	data science managers	dj patil, Marck Vaisman, Pete Skomoroch, John Myles White, Jon Bruner, Michael Dewar, Rika Gorn
Y03	democratic politicians	Nate Silver, Kamala Harris, Barack Obama, Joe Biden, Michelle Obama, Bill Gates, Vice President Kamala Harris
Y04	ecology/genetics faculty	Carl Zimmer, C. Brandon Ogbunu, Jeffrey Ross-Ibarra, Rasmus Nielsen, Jonathan Pritchard, Dmitri Petrov, Stephanie Spielman, PhD
Y05	generative artists	Mike Bostock, Susie Lu, Matt DesLauriers, zach lieberman, Kyle McDonald, Daniel Shiffman, The Pudding
Y06	HCI, visualization & graphics	Mike Bostock, Amanda Cox, Martin Wattenberg, Scott Murray, Fernanda Viégas, Tamara Munzner, Moritz Stefaner
Y07	jonathan haidt et al	Claire Lehmann, Douglas Murray, Sam Harris, Peter Boghossian, Jonathan Haidt, Maajid أبو عمّار, James Lindsay, getting one billion moms
Y08	open science	Michael Eisen, Ed Yong, Carly Strasser, Ethan White, jeremy freeman, Kaitlin Thaney 💁🏻‍♀️ (she/her), Open Science
Y09	open source devs	Josh Greenberg, timoreilly, Carly Strasser, Open Science, Leah Wasser 🦉offline thru early august, harper 🤯, Ben Marwick
Y10	python data tool devs	Peter Wang, Fernando Pérez, Wes McKinney, PyData, Jake VanderPlas, Andreas Mueller, Anaconda
Y11	python language devs	Guido van Rossum, PyCon US, Ewa Jodlowska, Nick Coghlan, Carol Willing, Brandon Rhodes, jacobian
Y12	python data scientists	Marc Hedlund, Andy Baio, Tyler Rinker, Sasha Laundy, Juliet Hof-Hu-How do you say Hougland?, Dr. Christie Bahlai, Frederick Solt
Y13	r devs from minority groups	Ayodele (eye-ya-deli) | Critical Bayes Theory, kaelen medeiros, Mine Dogucu, Maya Gans, Dr. Cat Hicks 📈👩🏻‍🔬🦄🏳️‍🌈, Cédric Scherer 💉, Daniela Vázquez
Y14	r devs not at rstudio	Andrie de Vries, Christophe Dervieux, timelyportfolio, Kirill Müller, Tareef Kawaf, Will Landau, Rich FitzJohn
Y15	rladies accounts	R-Ladies BuenosAires, R-Ladies Melbourne Inc, R-Ladies Istanbul, R-Ladies Madrid, R-Ladies DC, R-Ladies Munich, R-Ladies Nashville
Y16	rstudio tidyverse team	Hadley Wickham, Jenny Bryan, Mara Averick, David Smith, RStudio, Mine Çetinkaya-Rundel, Hilary Parker
Y17	tech ceos and vc	Elon Musk, Bill Gates, jack, Reid Hoffman, Patrick Collison, Guido van Rossum, timoreilly
Y18	tech critique/explainers	Leigh Honeywell, Alexandria Ocasio-Cortez, EricaJoy, Adrienne Porter Felt, Jessie Frazelle, bletchley punk, Lara Hogan
Y19	tech ethics	Arvind Narayanan, Carl T. Bergstrom, zeynep tufekci, Rumman Chowdhury, rediet abebe, Timnit Gebru, Safiya Umoja Noble PhD
Y20	tech thought leaders	Jonah Peretti, Andrew McLaughlin, steve o'grady, joshua schachter, John Lilly, brady forrest, Pete Warden

Saved data from Personalized PageRank random walks

See *full neighborhoods* of ~1,000 to ~10,000 nodes

and

partial neighborhoods of ~100,000 to ~10,000,000 nodes