Social network analysis of lexical diffusion

Investigating the spread of new words on Twitter

Quirin Würschinger

LMU München

ICAME 2019, Neuchâtel
Workshop 'Corpus approaches to social media'

1 May, 2019

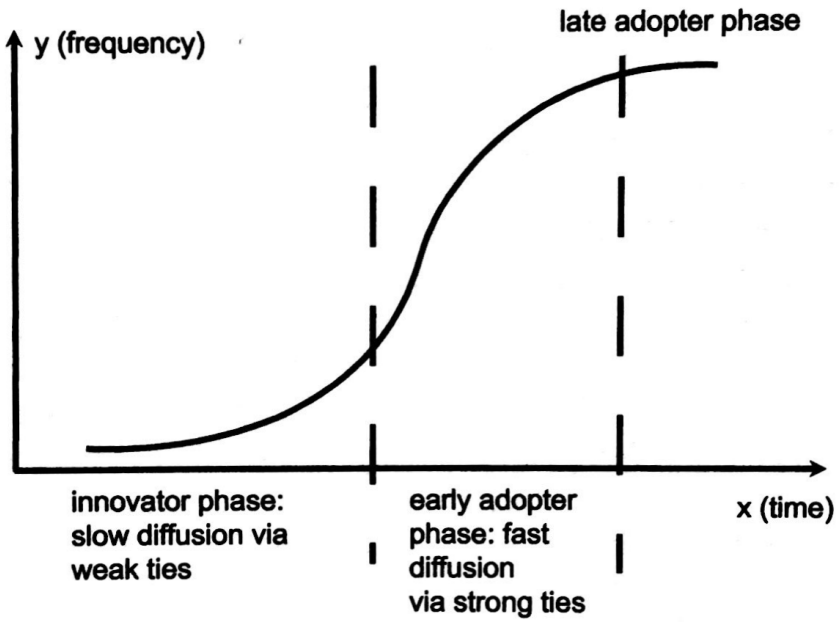
Outline

- Theoretical background
  - lexical innovation
  - diffusion
- Twitter approach
  - usage intensity
  - social network analysis
  - comparative analyses

What are 'lexical innovations'?

nonce formations	*ICAMErs*, *Neuchâtelgate*
↓	
neologisms	*microflat*, *burquini*, *bediquette*, *biobag*, *poppygate*, *emojinal*, *toxic*, *alt-right*, *bromance*, *Brexit*, *selfie*, *smartphone*
↓	
conventional lexemes	*phone*, *internet*, *laptop*

The S-curve model of diffusion



Early stages

- prestige of coiner and early adopters
- dense networks (J. Milroy and L. Milroy 1985)
- strong ties

Later stages

- diffusion via weak ties (Granovetter 1977)

How do new words diffuse?

Theoretical framework: The EC Model (Schmid forthc.)

Diffusion:

> "Linking the three aspects of speakers, cotexts, and contexts, I define diffusion as a process that brings about a change
>
> in the number of **speakers and communities** who conform to a regularity of co-semiotic behaviour and
>
> a change in the types of **cotexts and contexts** in which they conform to it."

Diffusion of lexical innovations

Research questions – two dimensions of diffusion:

1. How do neologisms diffuse across **usage contexts**?
   - increasing usage intensity (Stefanowitsch and Flach 2017)
   - increasing diversity in text types and semantic domains
2. How do neologisms diffuse across the **speech community**?
   - increasing number of speakers
   - increasing diversity of speaker communities

Diffusion across usage contexts

Previous empirical studies

- case studies: Hohenhaus 2006
- traditional corpora: Elsen 2004
- web corpora: Gérard 2017; Cartier 2017; Davies 2013; Kerremans, Stegmayr and Schmid 2012
- social media corpora: Grieve, Nini and Guo 2016

→ main focus: usage intensity

Diffusion across the speech community

Using Twitter data to study diffusion

- historical data
- informal and creative language use
- social media as a driving force in lexical innovation
- beyond usage frequency:
  - speaker information
  - sociolinguistic dynamics of diffusion

Collecting and processing Twitter data

Data collection

- Twitter's APIs
- scraping: `twint`

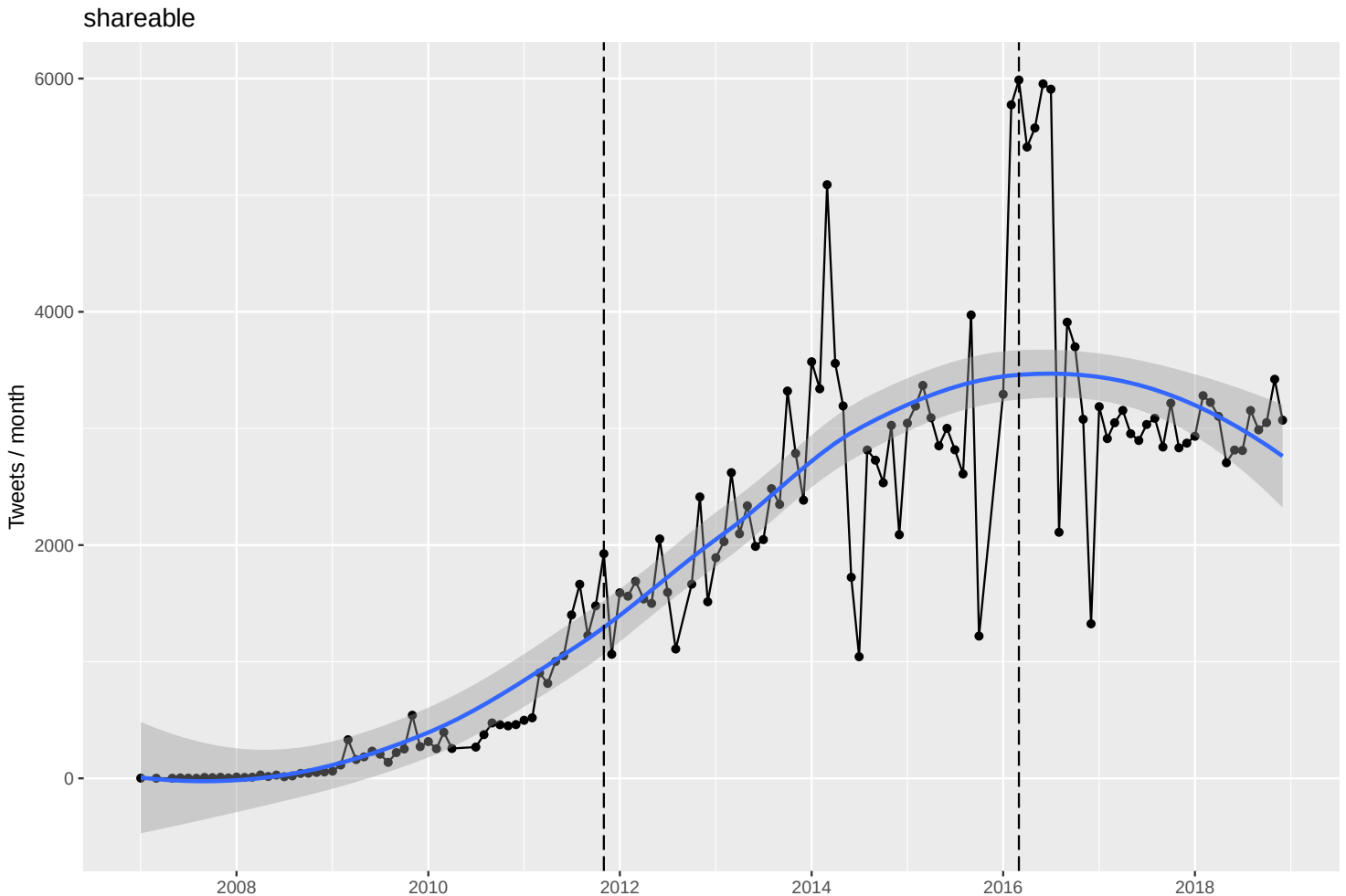
Data overview

- neologisms: 87
- timespan: 2006–2019
- number of tweets: 32 Mill.
- number of unique users: 13 Mill.

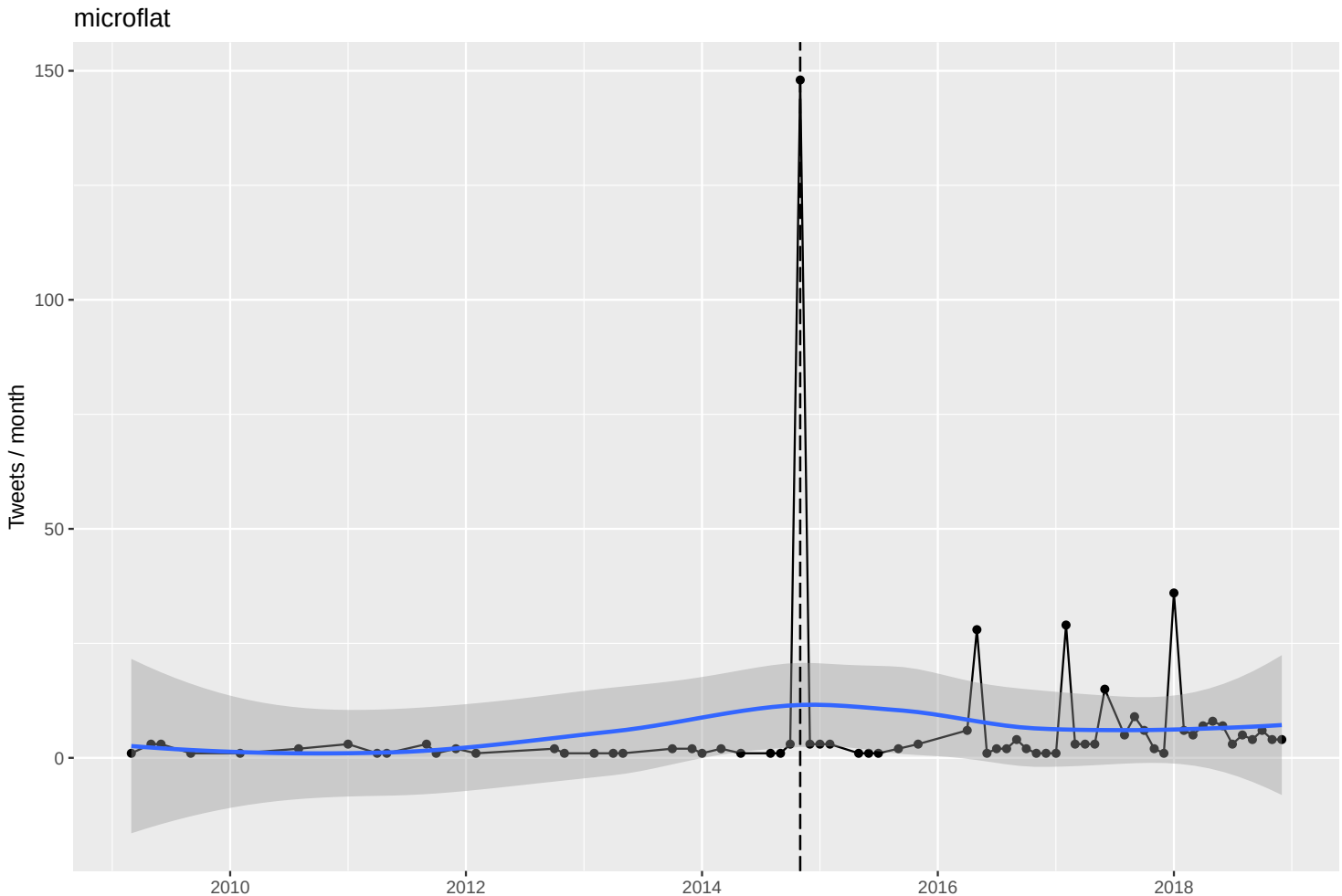
Focus of analysis

- degree of diffusion
  - usage intensity
  - social networks of diffusion
- diffusion stages

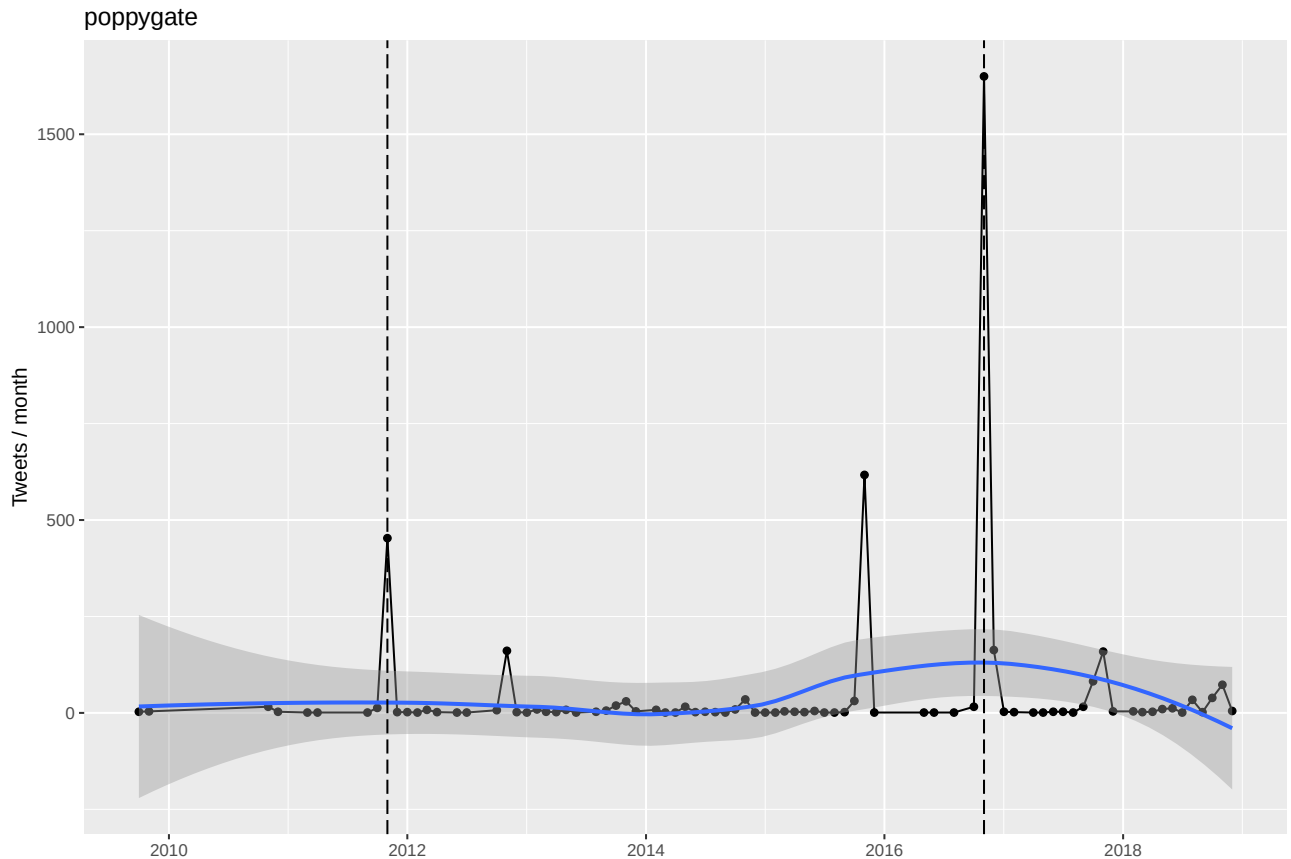
Advanced diffusion: *shareable*



Unsuccessful diffusion: *microflat*

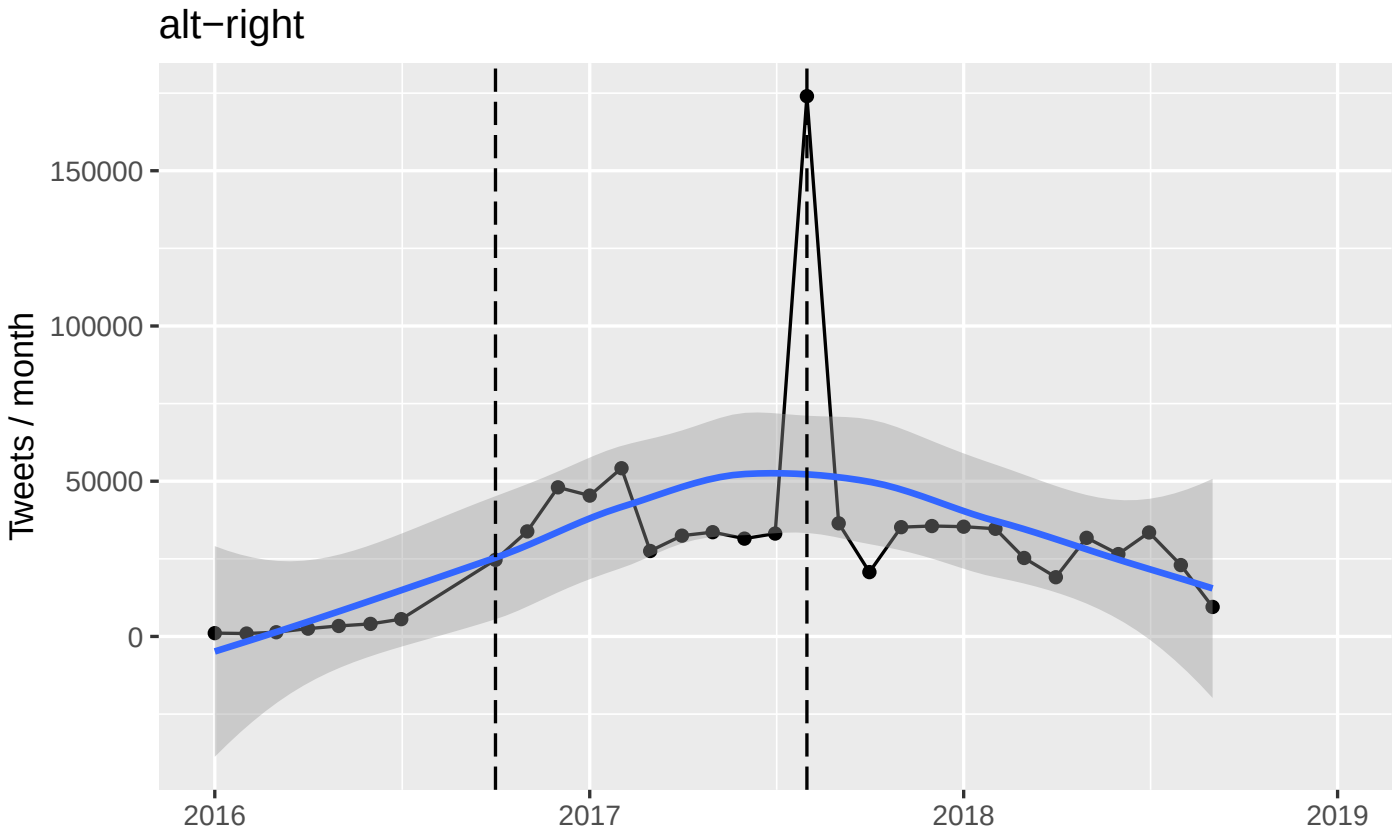


Topical diffusion: *poppygate*[1]



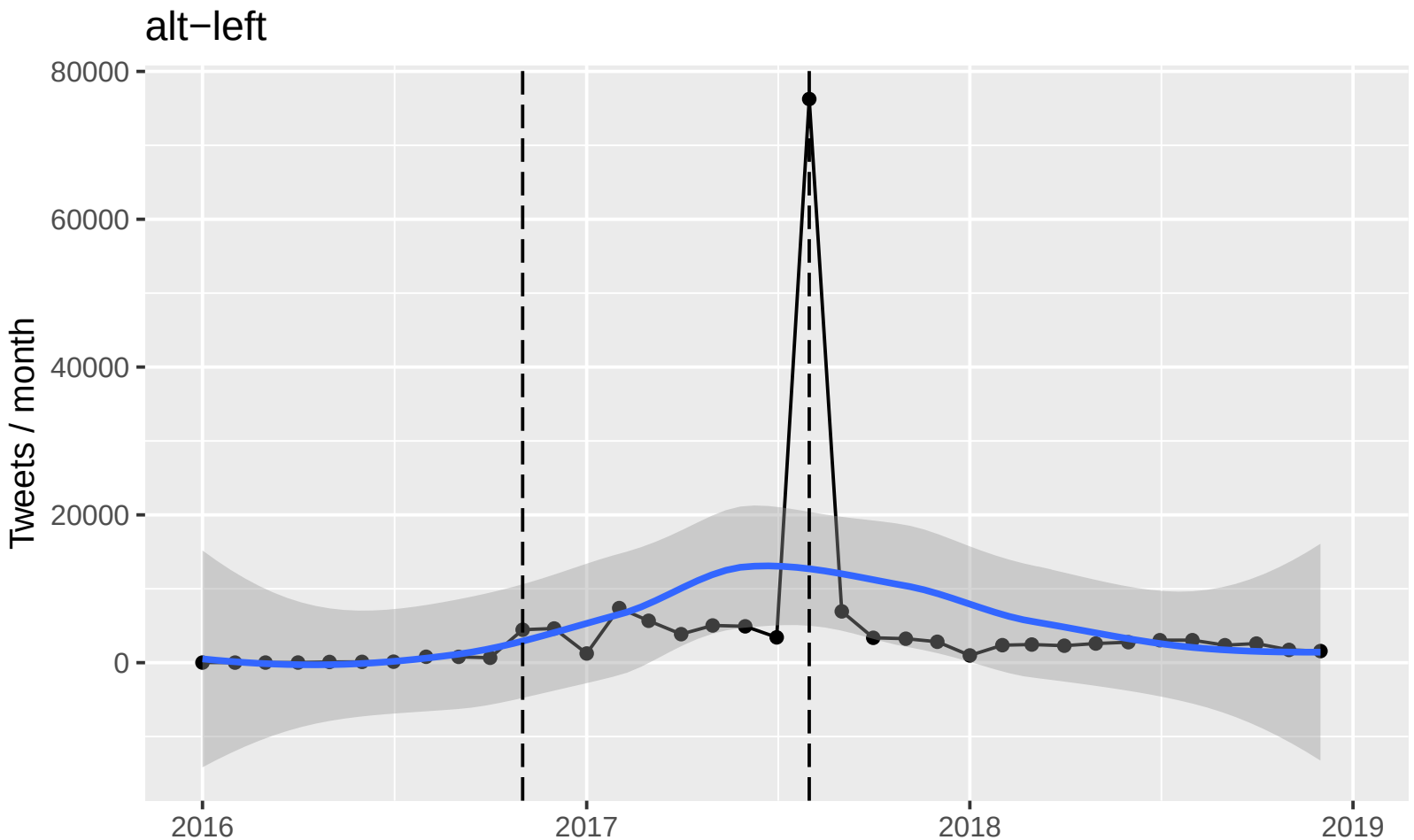
[1] *poppygate*: scandals around the ritual of wearing artificial flowers for Remembrance Day

Limited diffusion: *alt-right*[2]



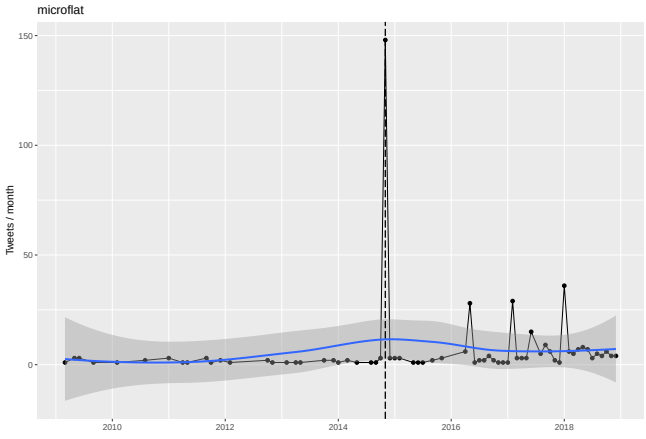
[2] *alt-right*: short for *Alternative Right* after White Supremacist Richard Spencer

Limited diffusion: *alt-left*

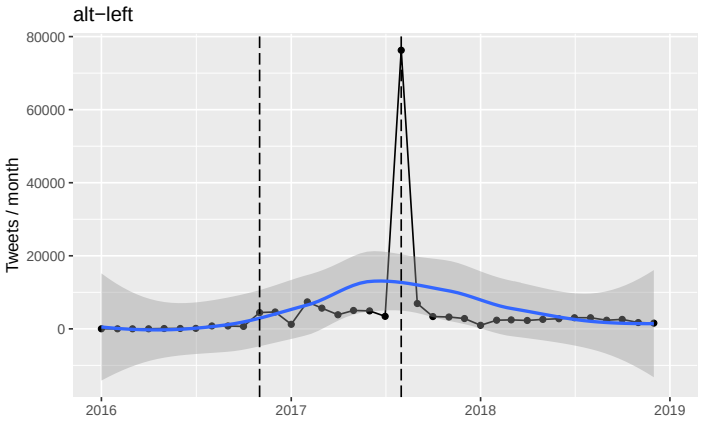


Degrees of diffusion – Clusters

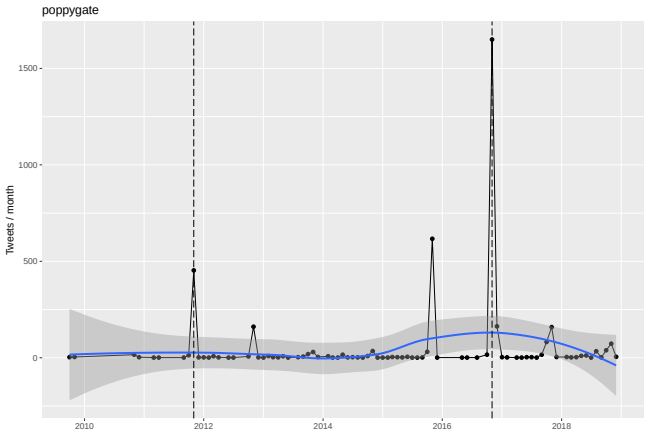
unsuccessful: *microflat*



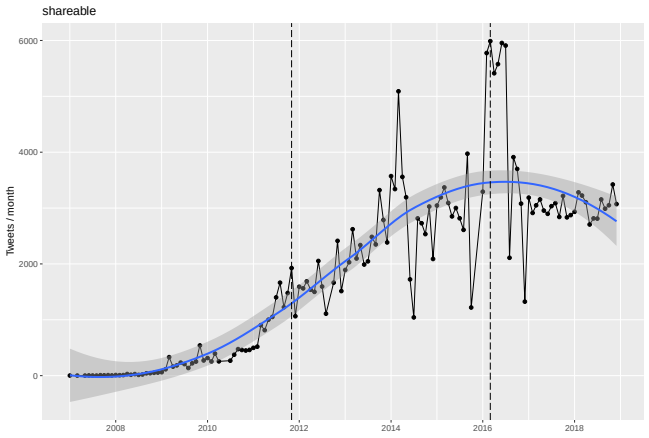
limited: *alt-left*



topical: *poppygate*



advanced: *shareable*



Corpus examples

use of *alt-left* in 2016

use of *alt-left* in 2017

Social network analysis[3]

Constructing the network

- extracting nodes and edges (`tidygraph`, `igraph`):
  - based on: mentions, retweets
  - data format: from text, from **columns** (`twint`)
- subsetting data: 1,000 interactions per
  1. first stage
  2. average usage intensity
  3. maximum usage intensity
  4. last stage

[3] all analyses and visualizations were done in `R`

Network structure[4]

Nodes

- node centrality: in-degree
- node positioning: Kamada-Kawai algorithm (`ggraph`)

Ties

- directionality: directed
- weights: degree

Communities

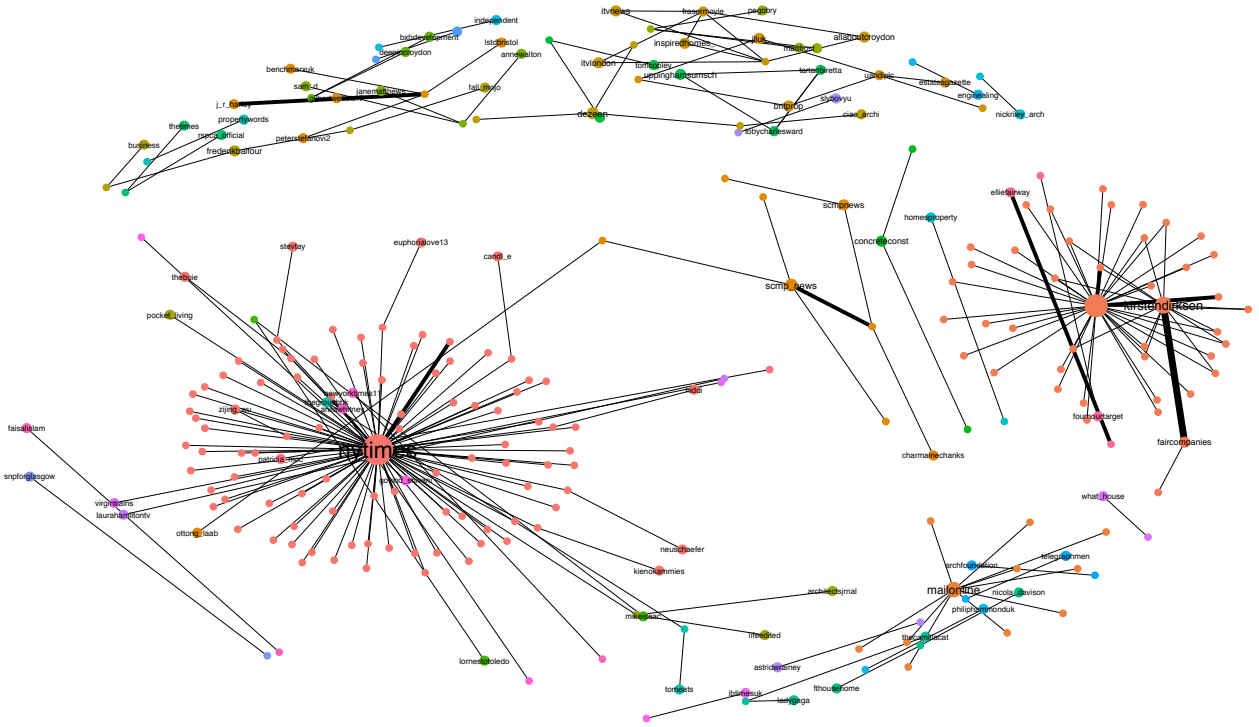
- clustering: edge betweenness algorithm
- modularity: fraction of ties within vs. between sub-communities

[4] all metrics and visualizations rely on `tidygraph` and `igraph`

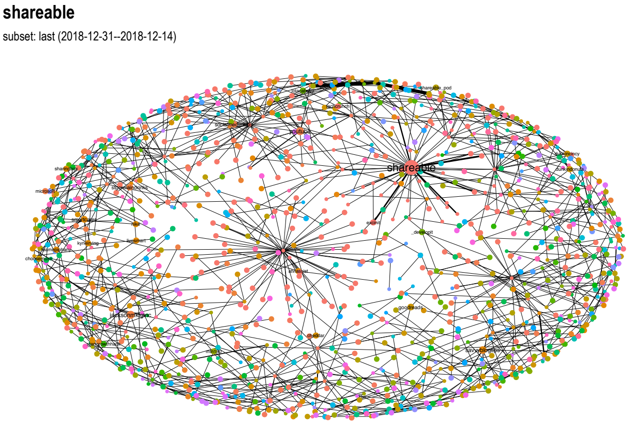
microflat

microflat

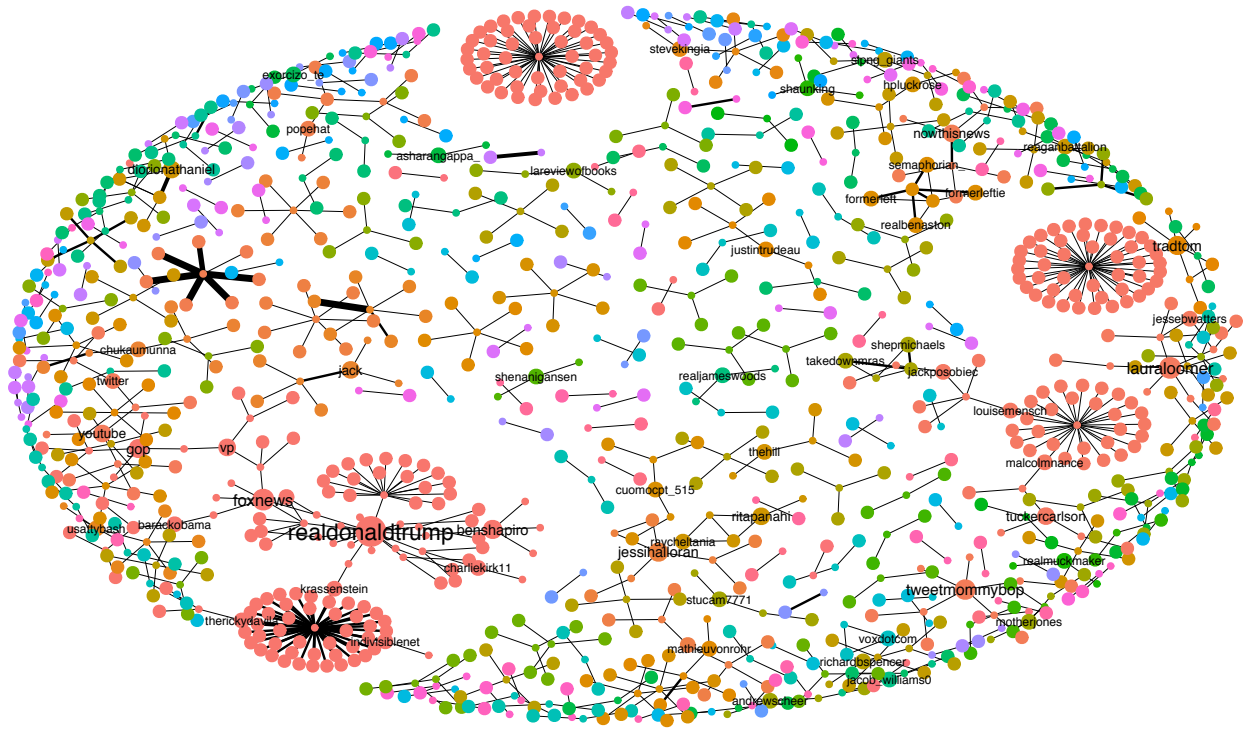
subset: last (2018-11-14--2011-01-06)



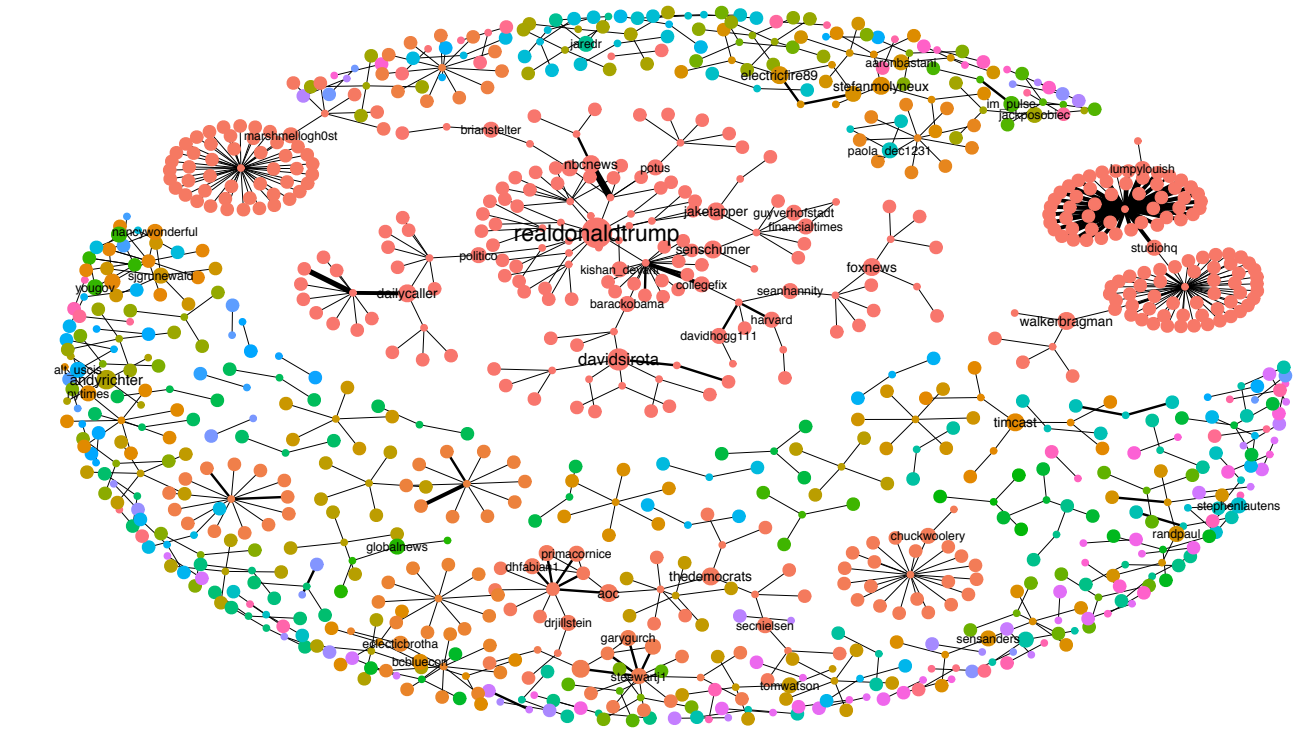
shareable



alt-right

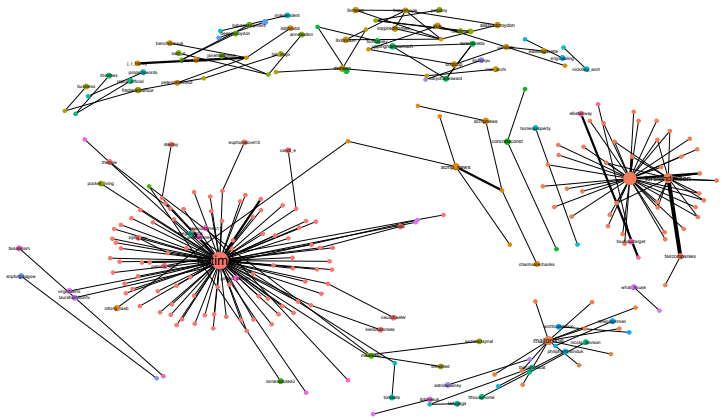

alt-right
subset: last (2018-09-10--2018-09-09)
stevekingia
exorcizo_te
popehat
asharangappa_
lareviewofbooks
oliodonathaniel
sheaunking
hpluckrose
nowthisnews
reaganbattalion
semanphorian_
formerleft
formerleftie
realbenaston
justintrudeau
tradtom
jessebwatters
chukaumunna
twitter
jack
shenanigansen
takedowmmras
shepmichaels
jackposobiec
realjameswoods
lauraloomer
louisemensch
youtube
gop
vp
thehill
malcolmnance
foxnews
cuomocpt_515
realdonaldtrump
benshapiro
ritapanahi
tuckercarlson
usattyhash
barackobama
raycheltania
jessihalloran
charliekirk11
realmuckmaker
krassenstein
stucam7771
tweetmommybop
motherjones
thenickydavis
individsiblenet
voxdotcom
mathieuvonrohr
richardbspencer
jacob_williams0
andrewsorkin

alt-left


alt-left
subset: last (2018-12-30--2018-12-19)
realdonaldtrump
nbcnews
potus
jaketapper
guyverhofstadt
financialtimes
senschumer
politico
kishan_devani
collegefix
seanhannity
foxnews
dailycaller
barackobama
harvard
davidhogg111
davidsirota
jaredr
aaronbastani
electricfire89
stefanmolyneux
im_pulse
jacksposobiec
paola_dec1231
marshmellogh0st
brianstelter
lumpylouish
studiohq
walkerbragman
nancywonderful
signpnewald
yougov
andyrichter
nytimes
alt_usnpk
timcast
stephenlautens
randpaul
chuckwoolery
globalnews
primacornice
dhfabian1
aoc
thedemocrats
drjillstein
secnielsen
garygurch
eclectibrotha
bcbluecgn
steewerj1
tomwatson
sensanders

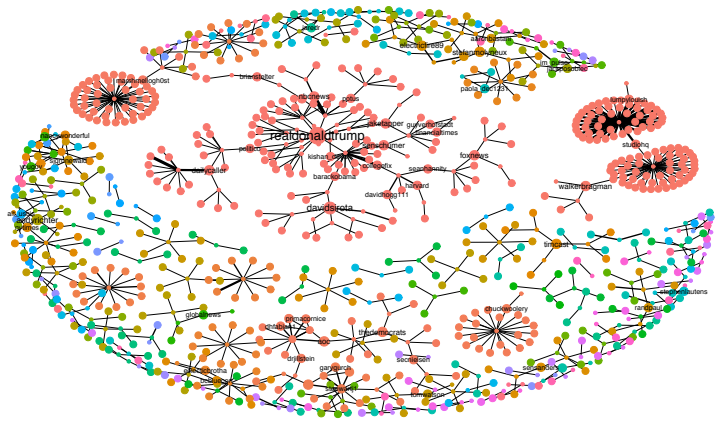
microflat

subset: last (2018-11-14--2011-01-06)



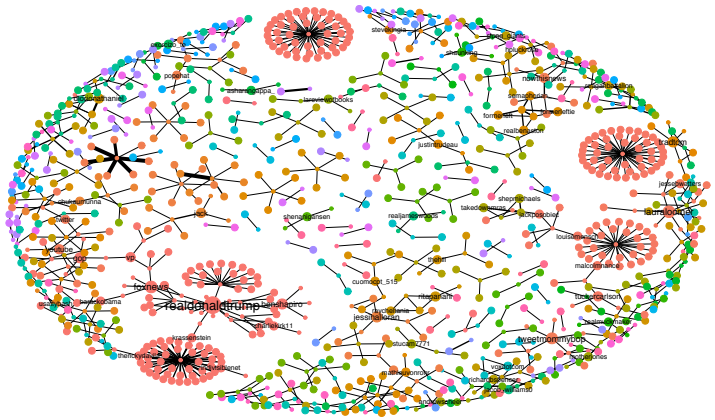
alt-left

subset: last (2018-12-30--2018-12-19)



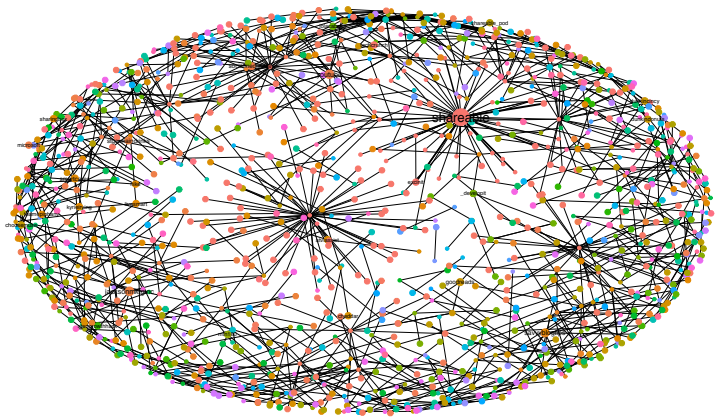
alt-right

subset: last (2018-09-10--2018-09-09)

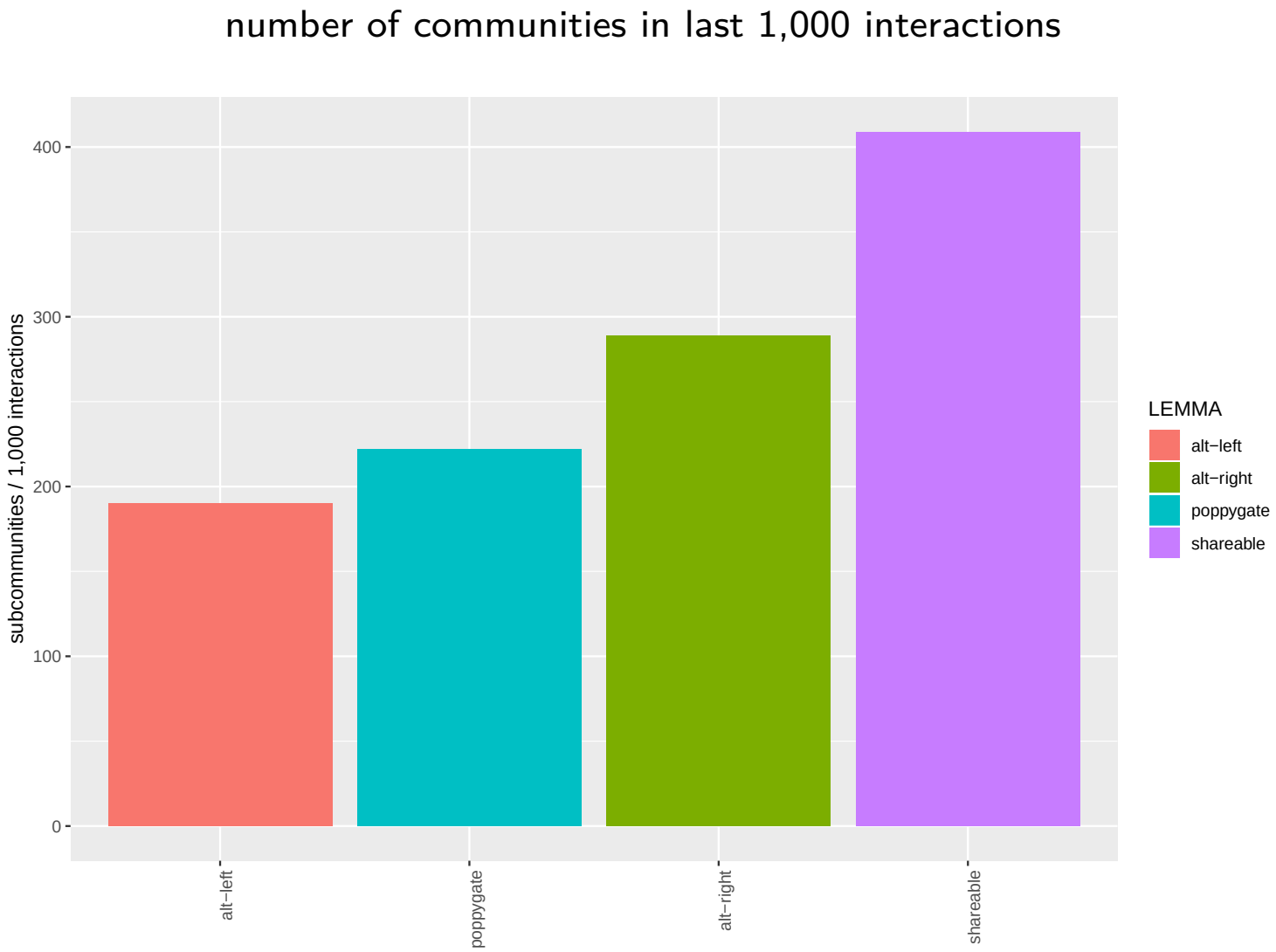


shareable

subset: last (2018-12-31--2018-12-14)



Comparison: metrics

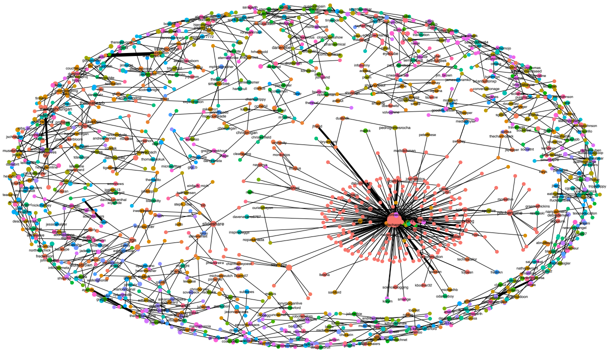


Stages of diffusion

First stage

shareable

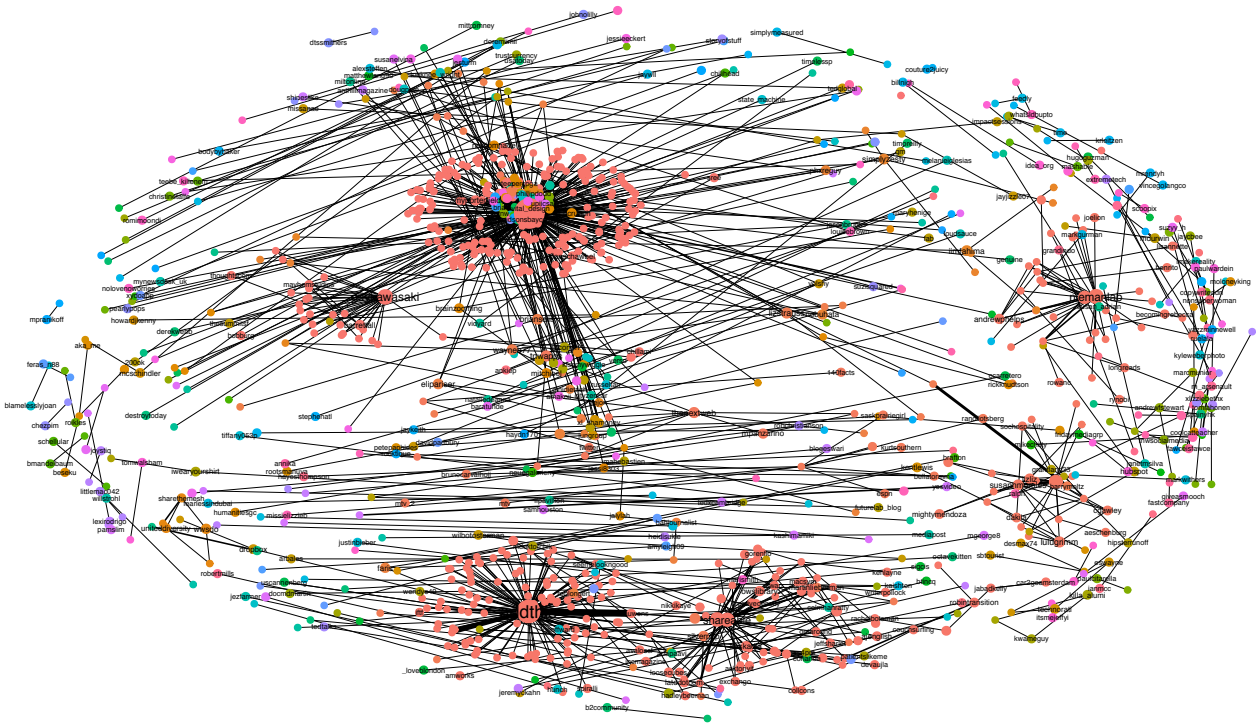
subset: first (2007-07-19--2009-08-02)



Second stage

shareable

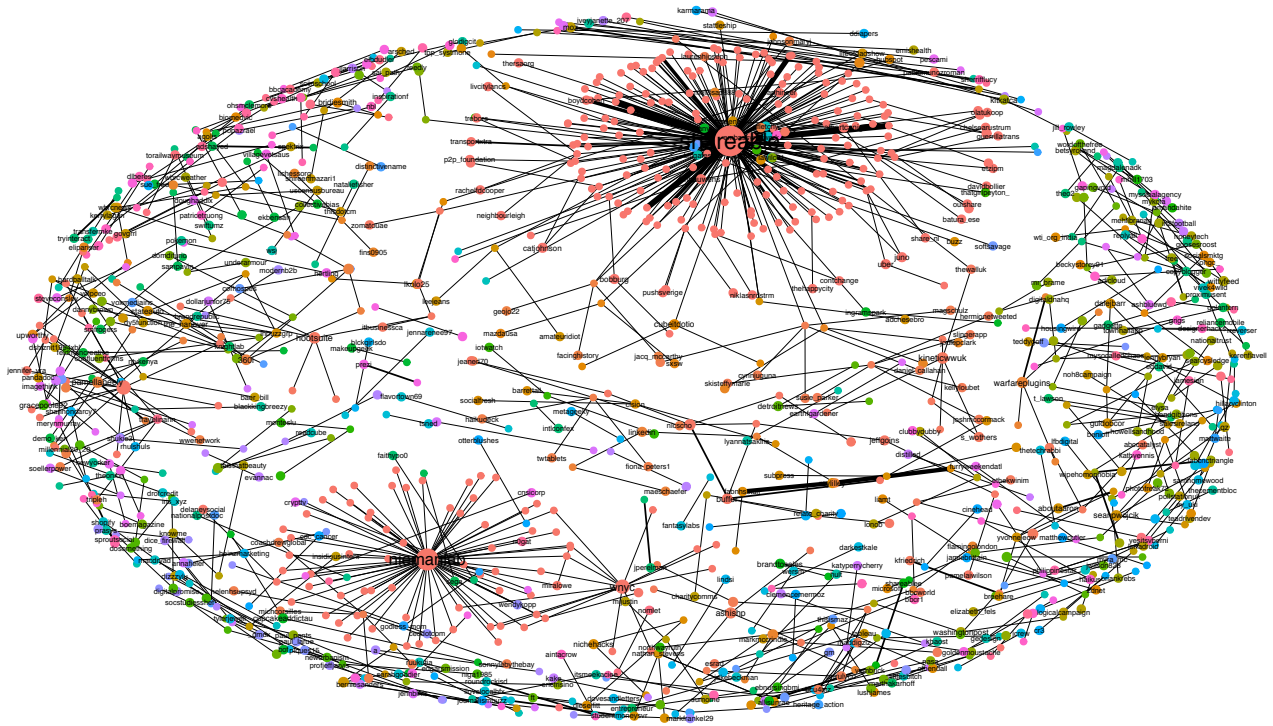
subset: mean (2011-11-02--2011-11-28)



Third stage

shareable

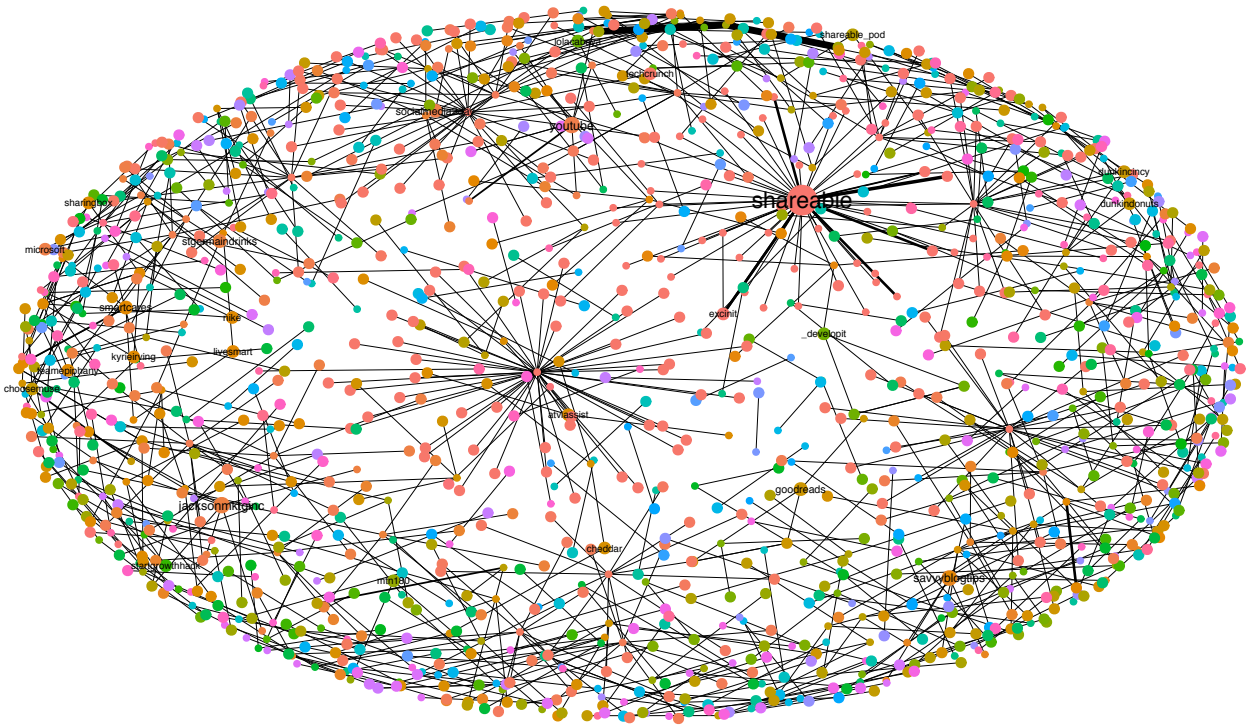
subset: max (2016-03-02--2016-03-18)



Fourth stage

shareable

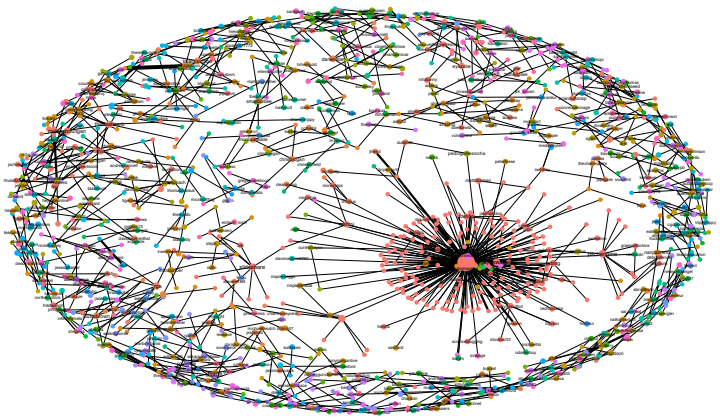
subset: last (2018-12-31--2018-12-14)



All stages

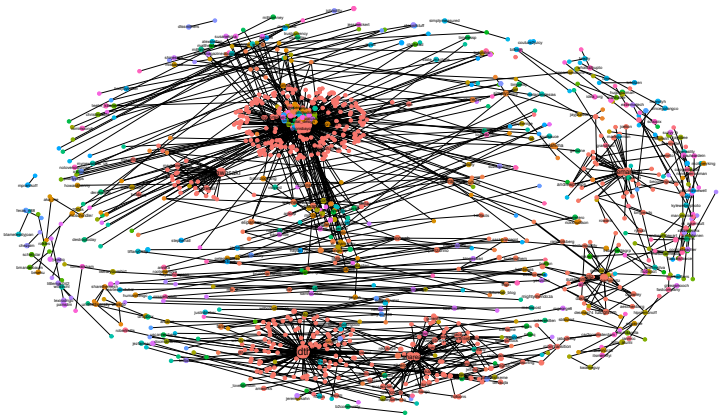
shareable

subset: first (2007-07-19--2009-08-02)



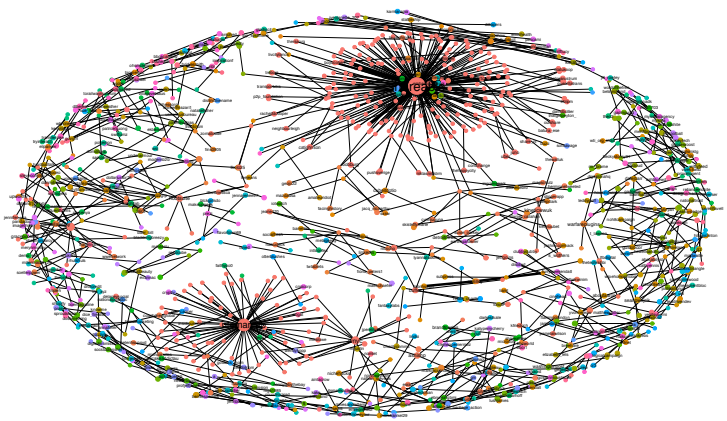
shareable

subset: mean (2011-11-02--2011-11-28)



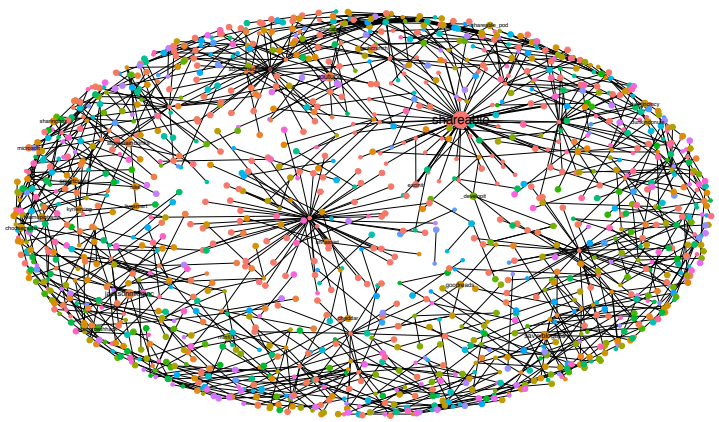
shareable

subset: max (2016-03-02--2016-03-18)

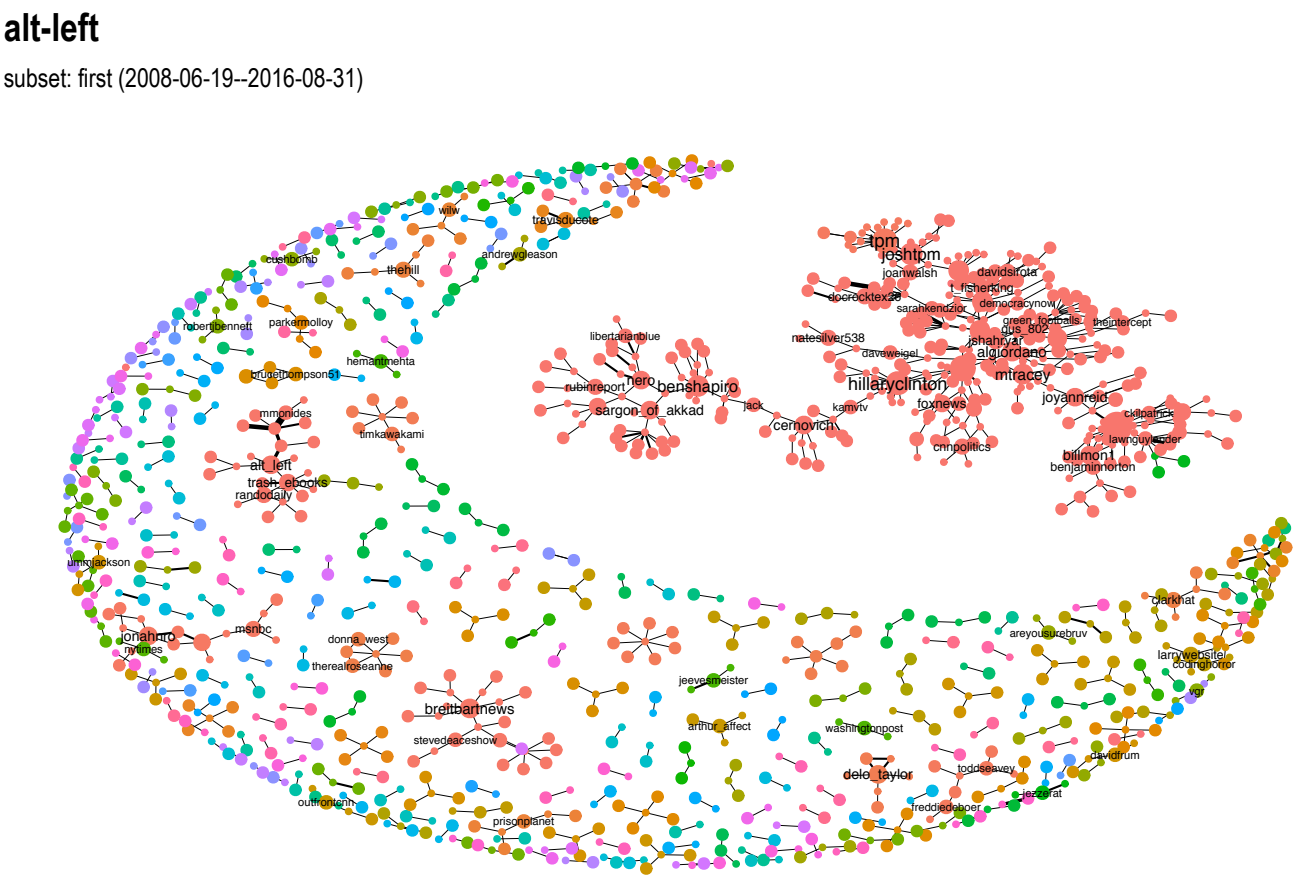


shareable

subset: last (2018-12-31--2018-12-14)



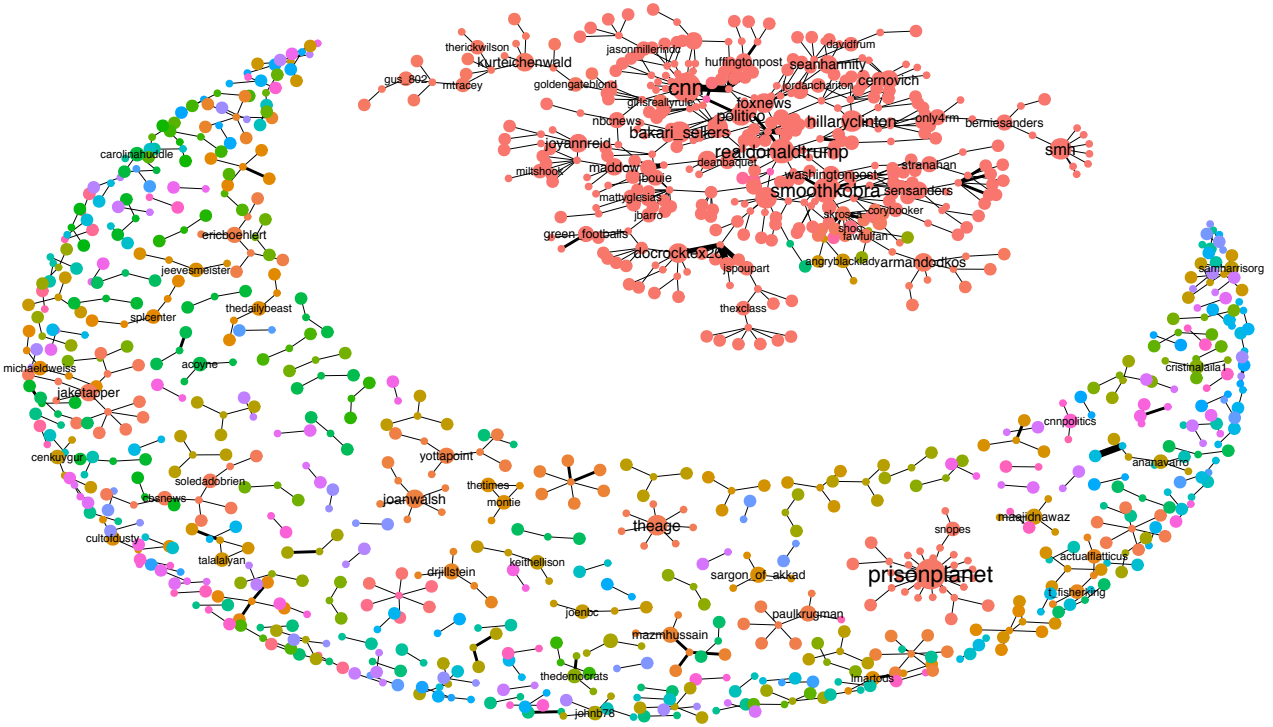
First stage


alt-left
subset: first (2008-06-19--2016-08-31)
wilw
travisdbeetle
andrewjpleason
cashbonib
thehill
robertbennett
parkermolloy
hemantmehta
brucethompson51
libertarianblue
rubinreport
nero
benshapiro
sargon_of_akkad
jack
cernovich
natesilver538
daveweigel
hillaryclinton
foxnews
kamvtv
cnnpolitics
mtracey
ipm
joshtpm
joanwalsh
davidsirota
t_fisherking
docrocktex38
sarahkendzior
democracynow
green_football5
theintercept
dug_802
ishahriyar
algiordano
joyannreid
billmon1
benjaminnorton
clilpatrick
lawnguyleader
mmmmides
timkawakami
alt_left
trash_ebooks
randomdaily
timmjackson
jonahnro
nytimes
msnbc
donna_west
therealroseanne
breitbartnews
stevedeaceshow
outfrontcnn
prisonplanet
jeevesmeister
arthur_affect
washingtonpost
dele_taylor
toddseavey
freddiedeboer
jezzorat
davidfrum
areyousurebruv
clarkhat
larrywebsite
coalitionforr
wsj

Second stage

alt-left

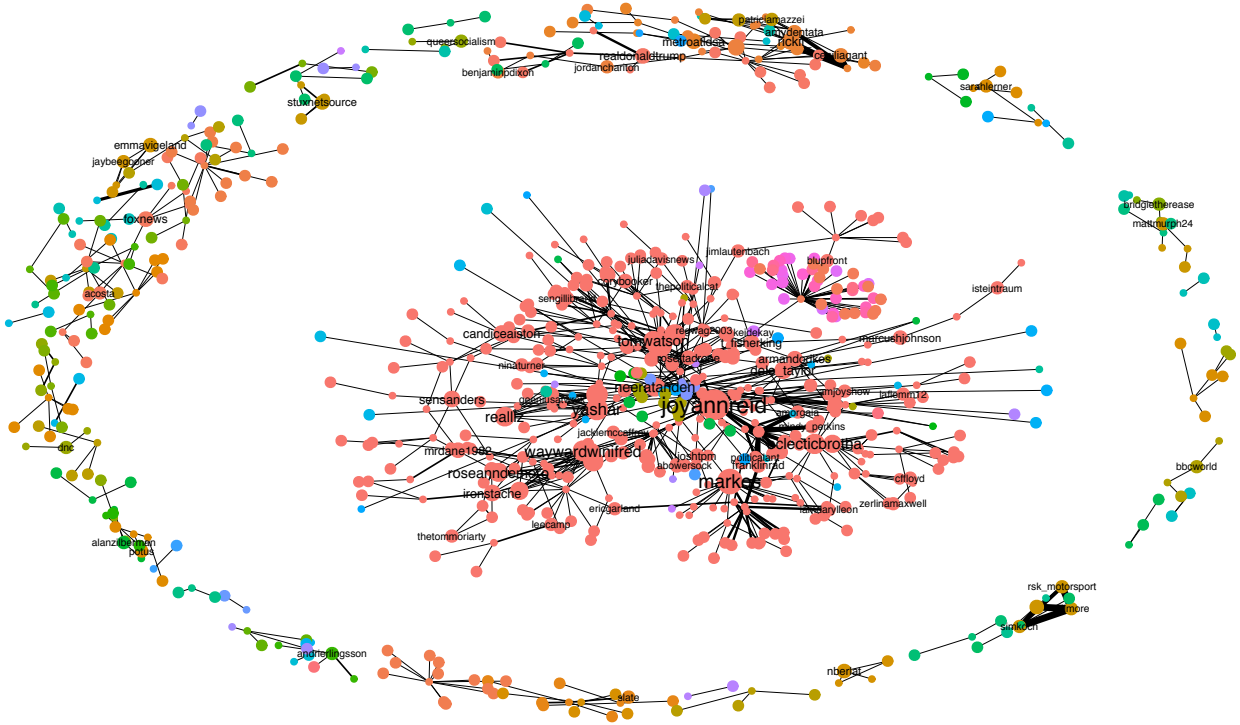
subset: mean (2016-11-02--2016-11-15)



Third stage

alt-left

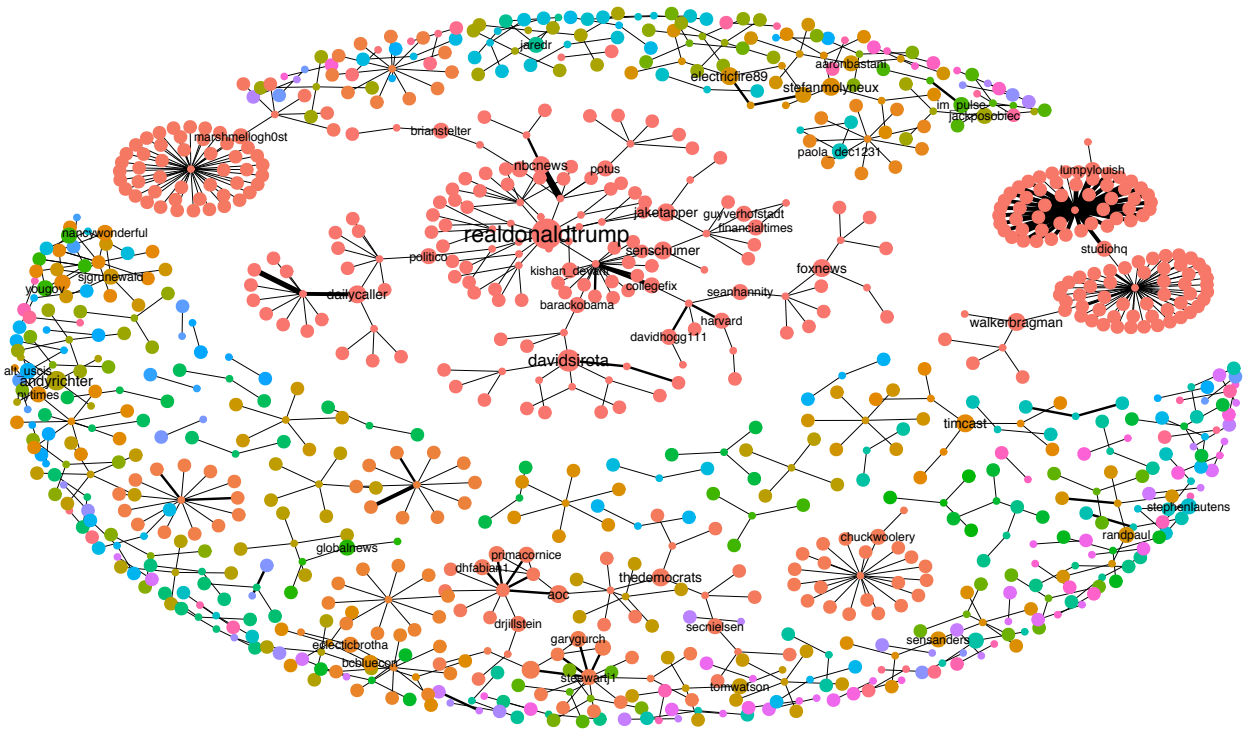
subset: max (2017-08-02--2017-08-03)



Fourth stage

alt-left

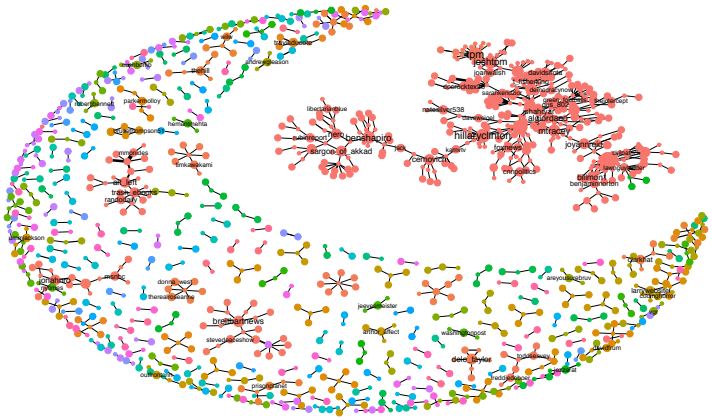
subset: last (2018-12-30--2018-12-19)



All stages

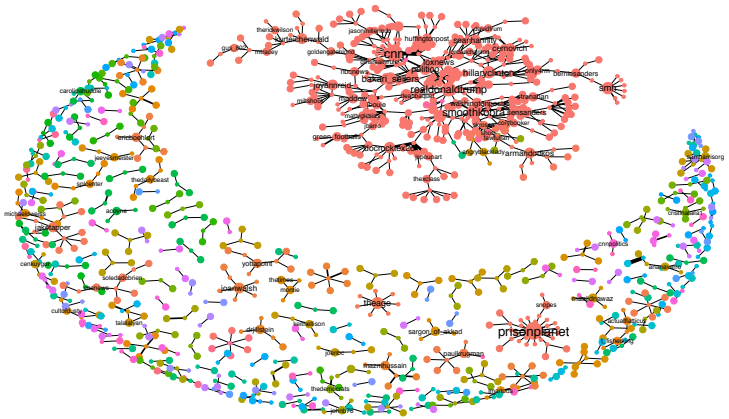
alt-left

subset: first (2008-06-19--2016-08-31)



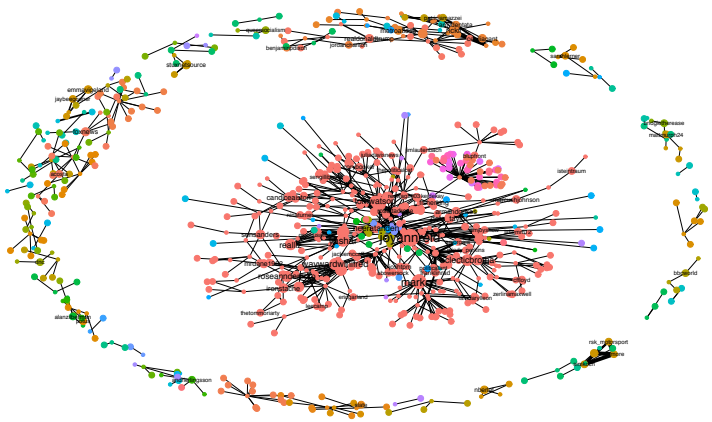
alt-left

subset: mean (2016-11-02--2016-11-15)



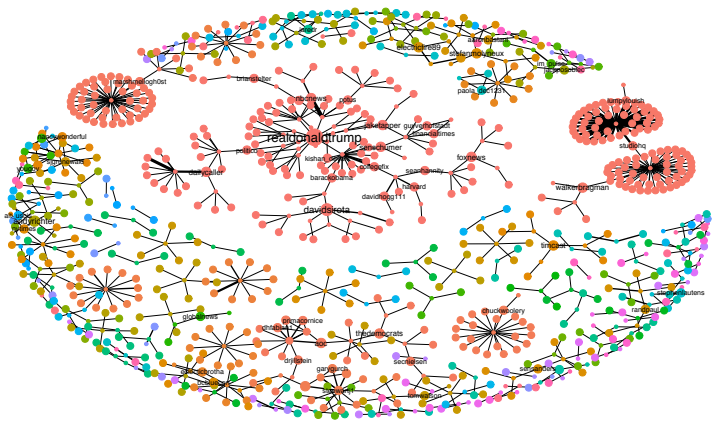
alt-left

subset: max (2017-08-02--2017-08-03)

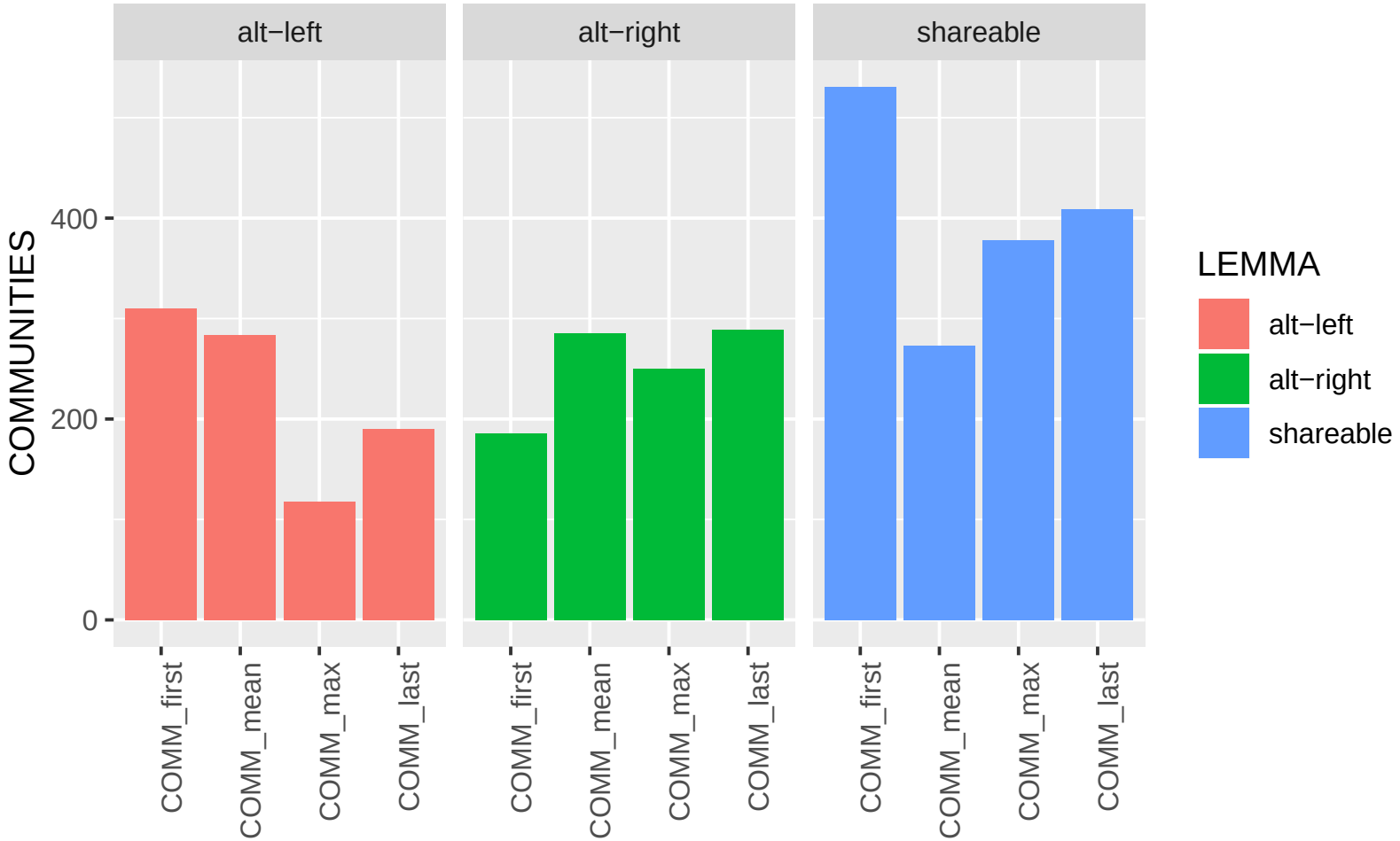


alt-left

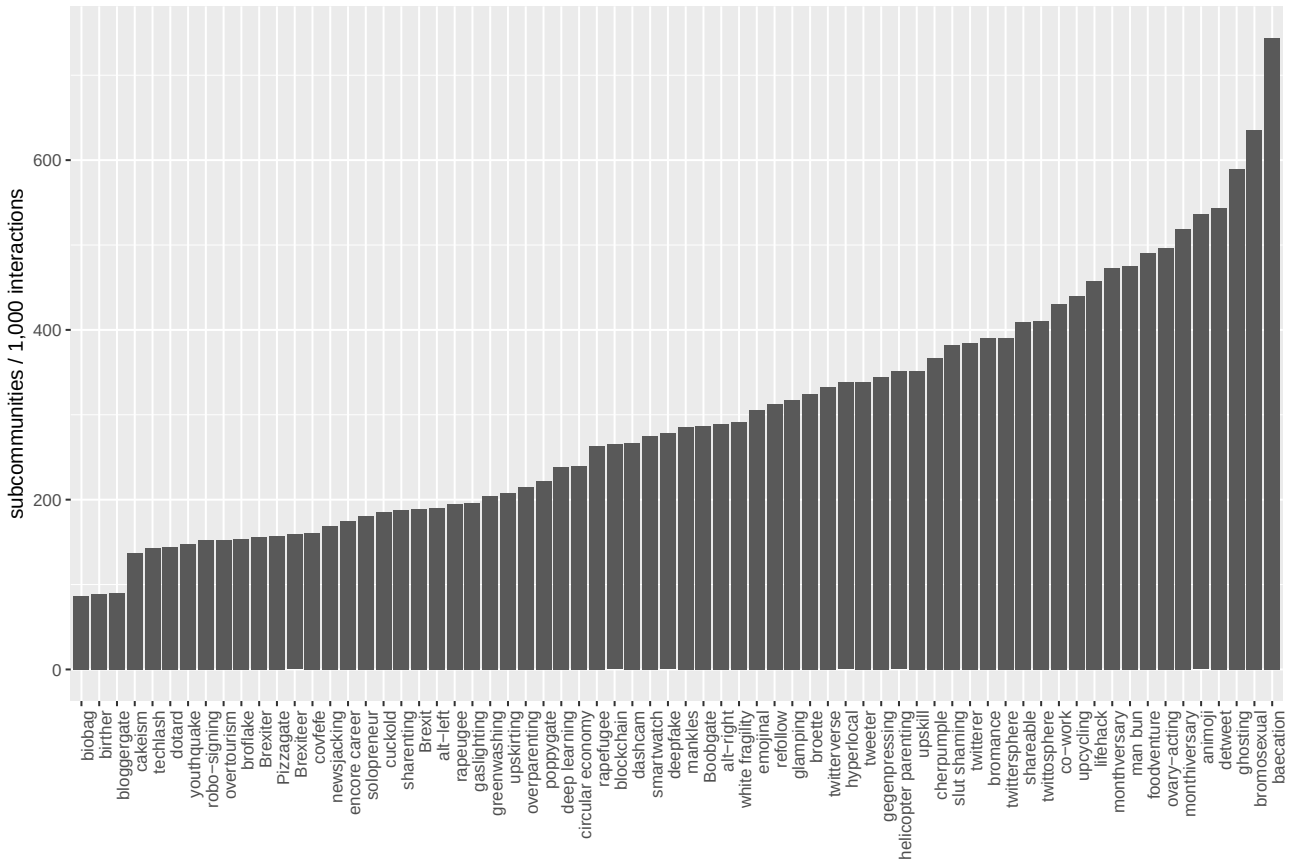
subset: last (2018-12-30--2018-12-19)



Comparison of diffusion stages across lexemes



Comparison of communities across all lexemes

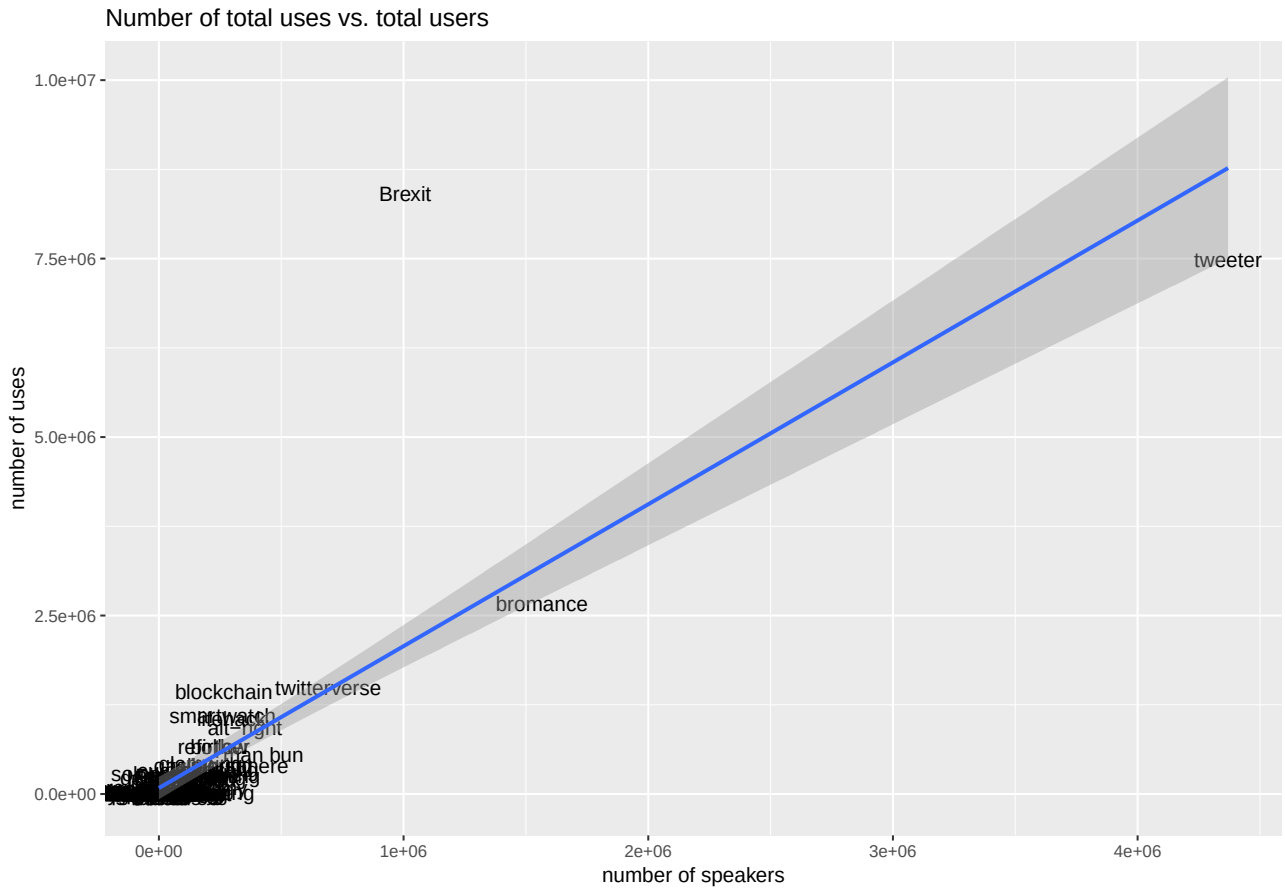


e.g. *covfefe*, *dotard*, *birther*, *Pizzagate*

e.g. *animoji*, *detweet*, *man bun*, *monthversary*

Users vs. usage

e.g. *Brexit*, *blockchain*, *smartwatch*



e.g. *tweeter*, *bromance*, *man bun*, *ghosting*

Conclusion: zooming out again ...

- Social media data make it possible to go beyond usage intensity to study the sociolinguistic dynamics of diffusion.
- This is particularly important for cases of limited social diffusion.
- Dense networks promote diffusion in earlier stages of diffusion.
- Weak ties are crucial for advanced diffusion to new parts of the speech community at later stages.

Discussion

Legal issues

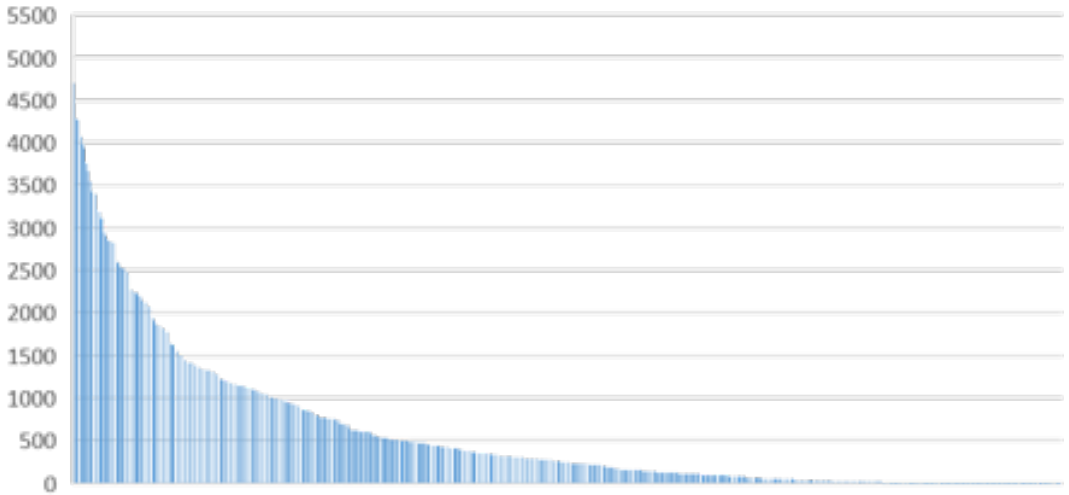
- scraping
- data vendors
- Twitter authorization
- user privacy

Thanks for your attention!

The NeoCrawler

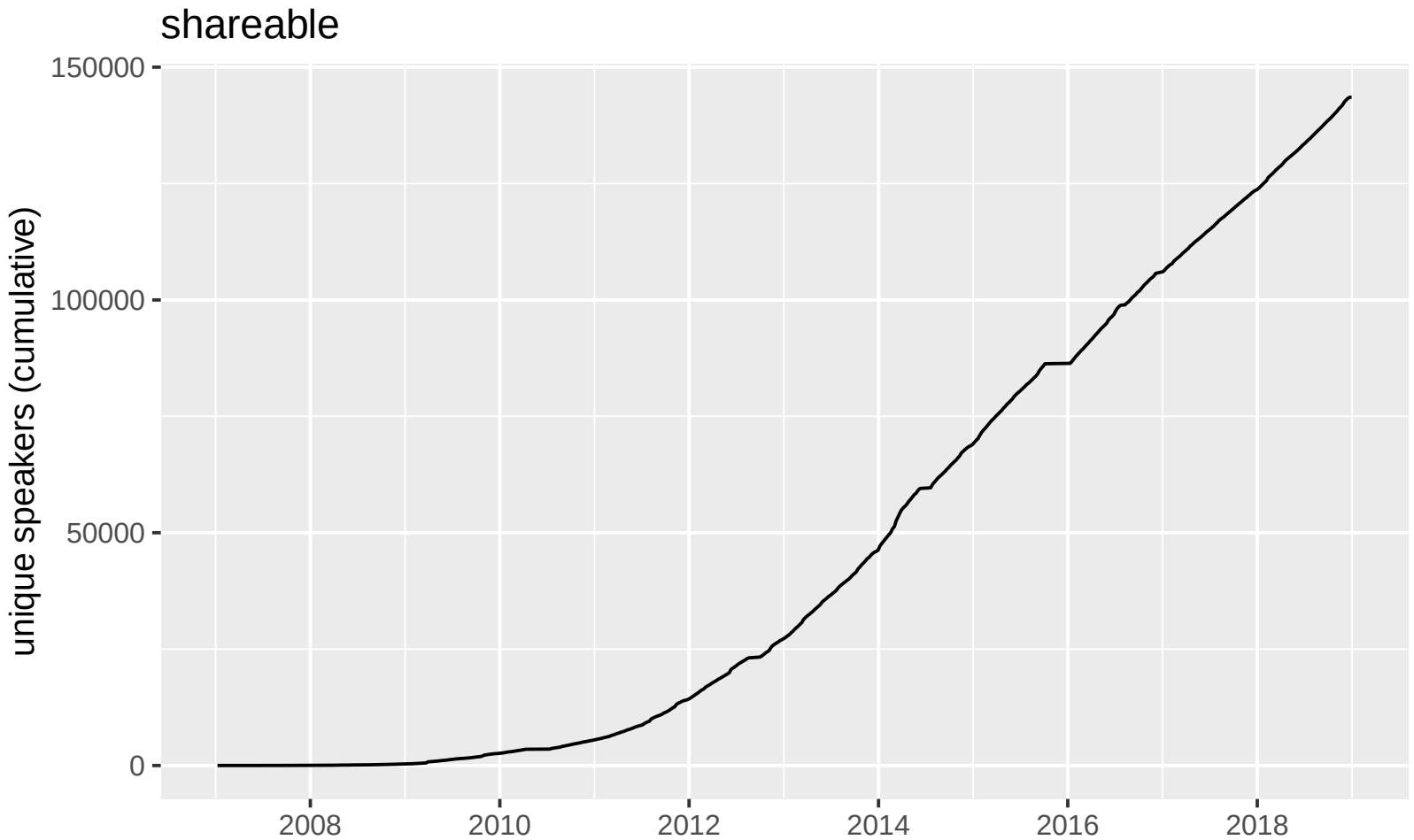
(Kerremans, Stegmayr and Schmid 2012)

Monitoring diffusion *across usage contexts* on the WWW

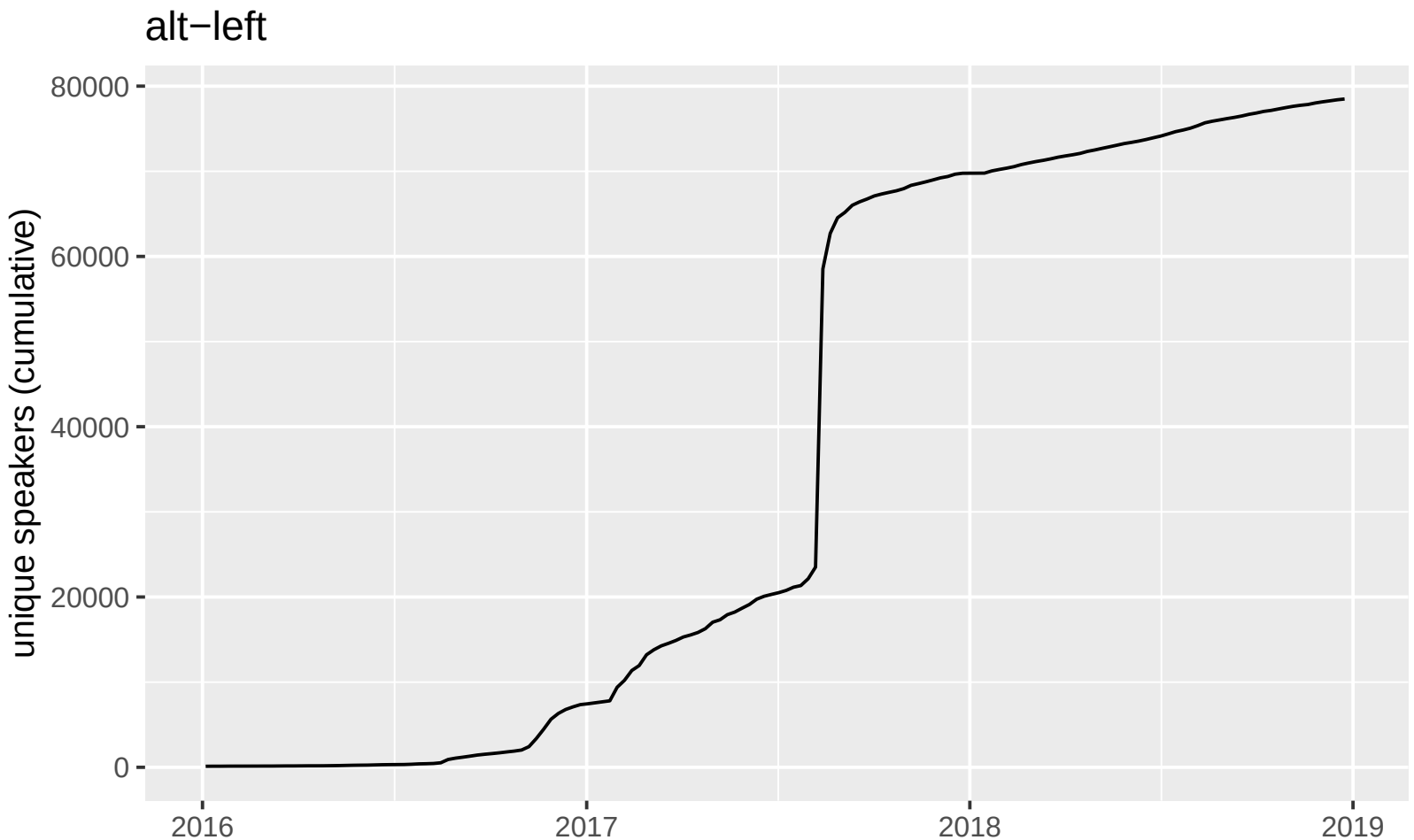


- sample: ≈ 1,000 candidates
- time window: 2011–2018
- corpus: ≈ 800,000 pages
- usage contexts: private forums, blogs, newspaper websites etc.

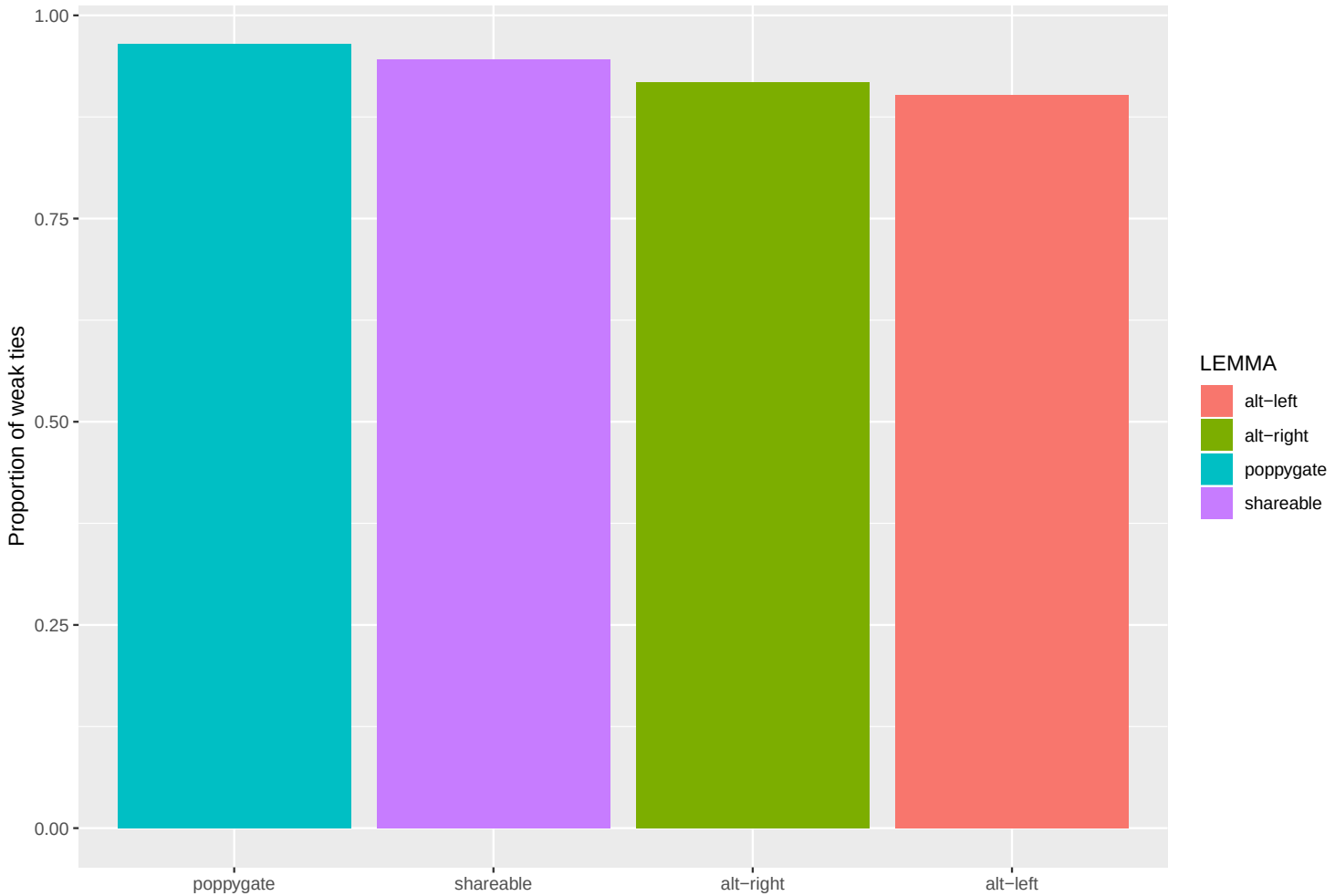
Increase in speakers for *shareable*



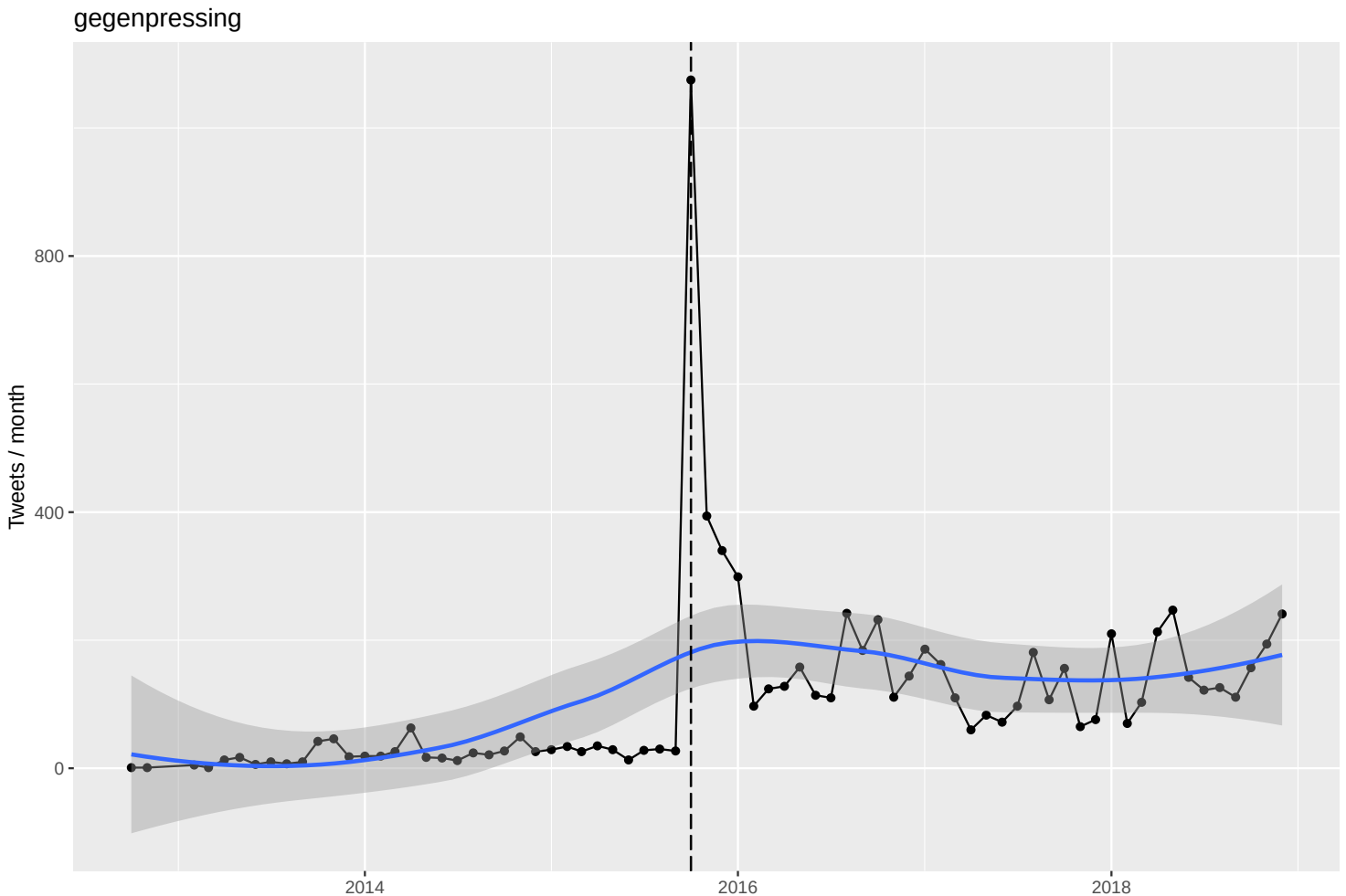
Increase in speakers for *alt-left*



The role of weak ties



Diffusion of *gegenpressing*



Social network of *gegenpressing*