Category Diffusion in the Reddit Network

Robert Johns

Overview of Reddit

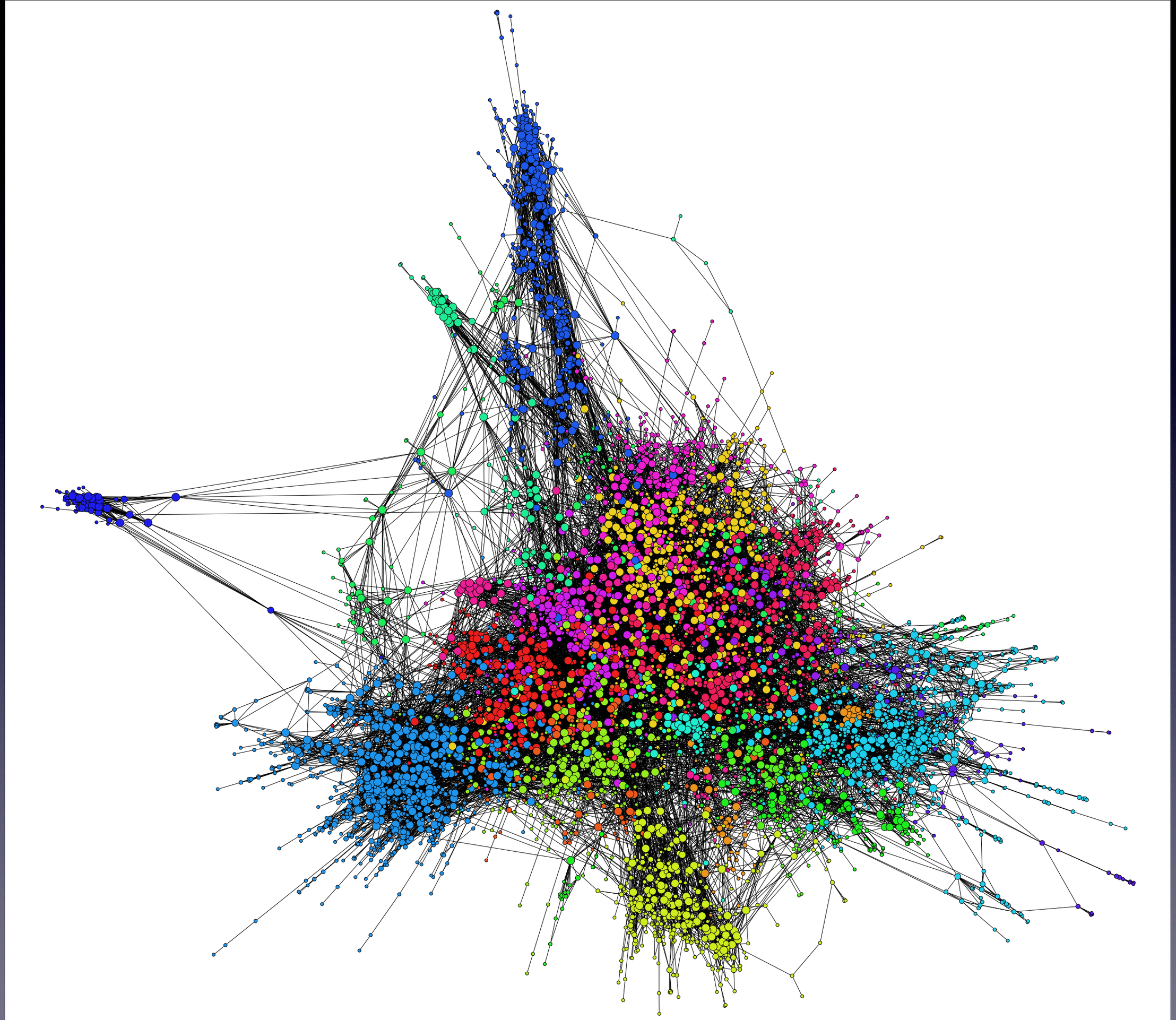
- Collection of communities, called subreddits (> 450k, ~10k active)
- Each subreddit is devoted to a topic:
  - baseball
  - pictures of dogs
  - reddit
- Front page of a subreddit consists of
  - user posts (links, discussion)
  - sidebar with links to related subreddits

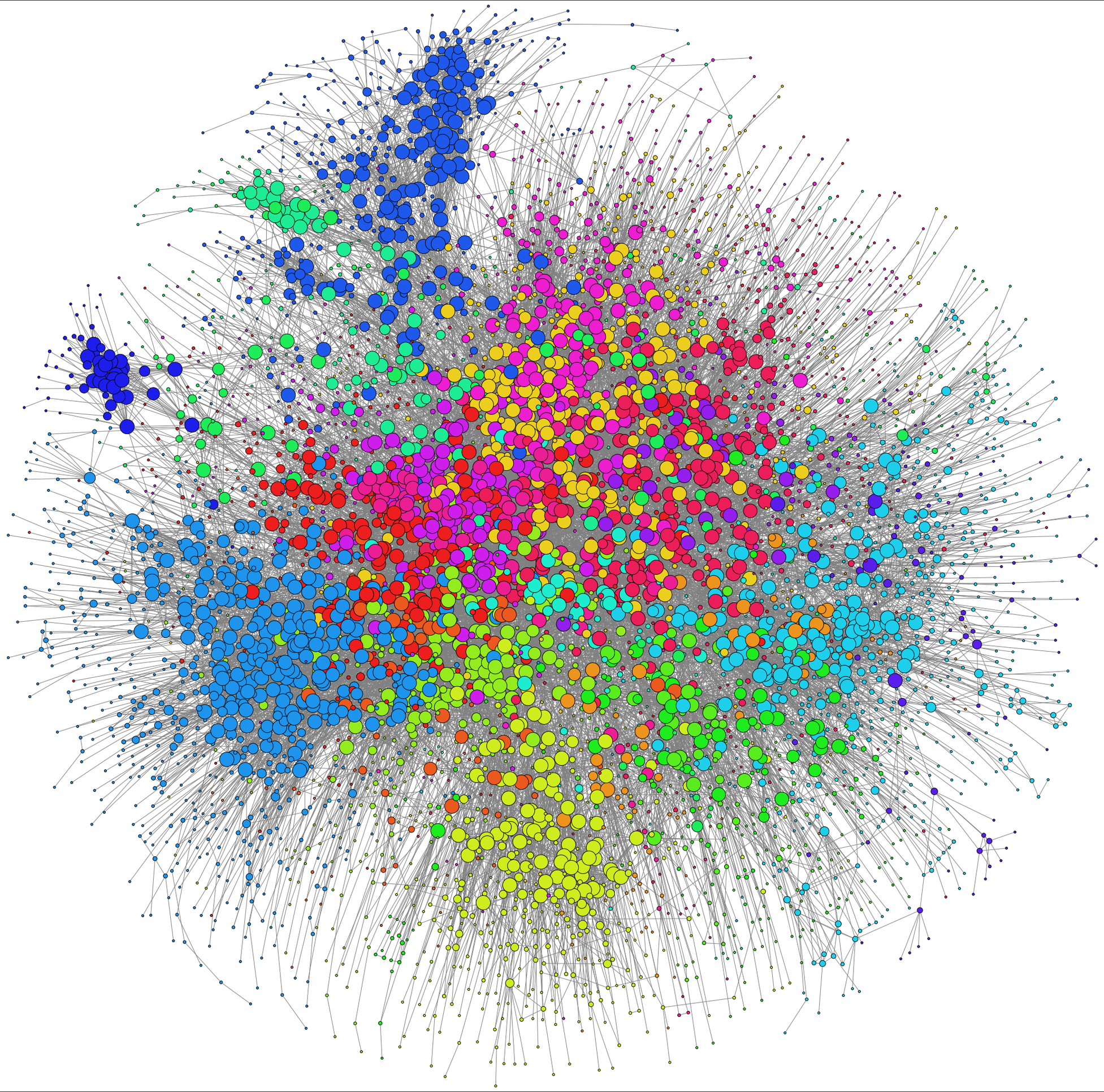
Idea & Hypothesis

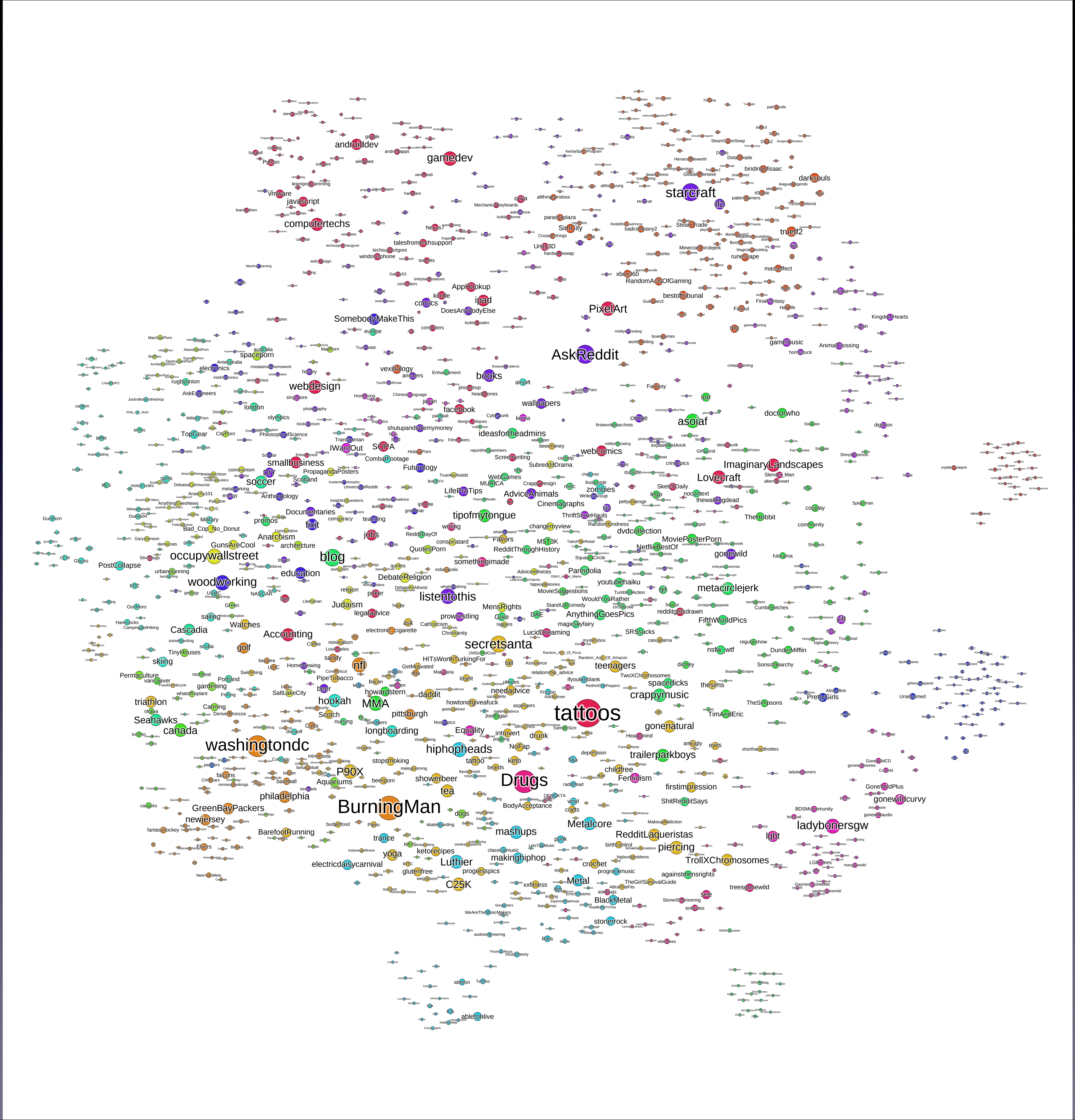
- Reddit as a graph
  - subreddits -> nodes
  - sidebar links -> edges
- That way, we could study the structure of reddit as a community of communities

Data Collection

- First try
  - get list of top 5000 subreddits by size from external website
  - crawl through html of each subreddit to get related subreddits, upload to csv or database
    - problem: reddit's API
- Second try
  - research paper on Github with similar topic
  - 5,288 nodes
  - 22,400 edges

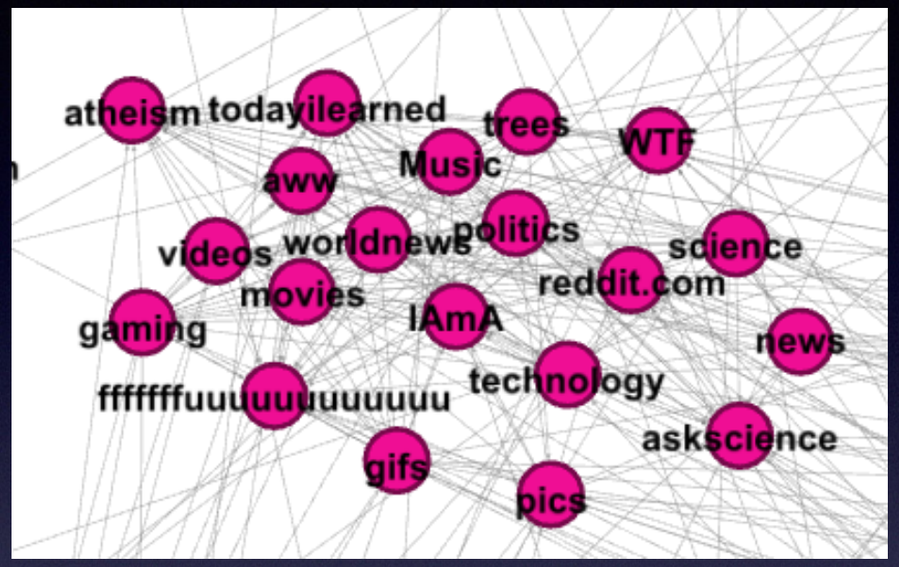


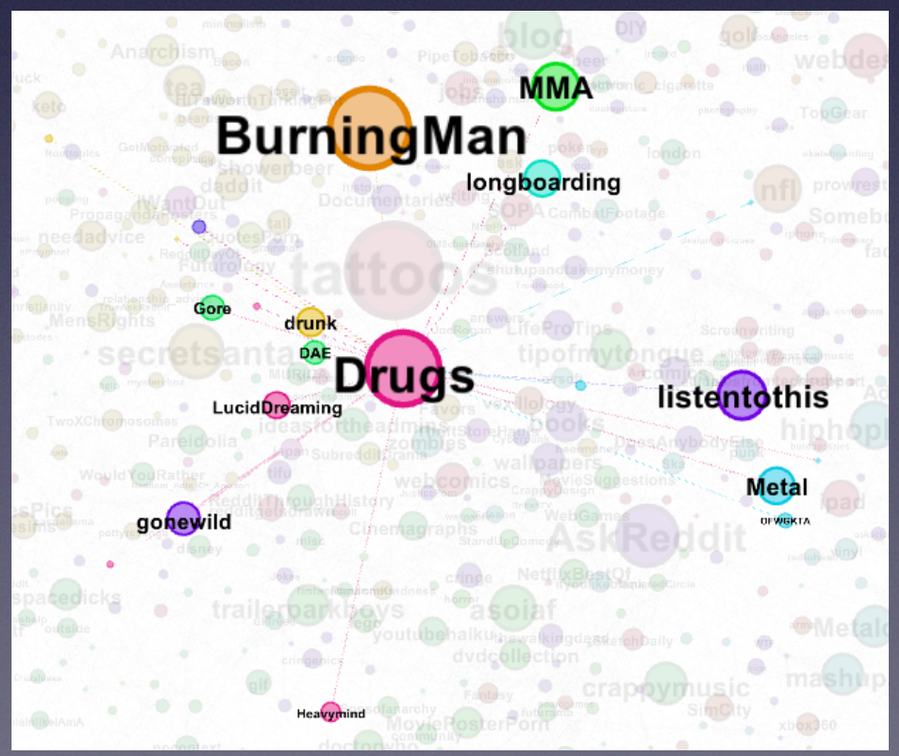



androiddev
gamedev
starcraft
darksouls
vmware
javascript
computertechs
tf2
truetf2
talesfromtechsupport
SimCity
runescape
masseffect
AppleHookup
ipad
comics
PixelArt
SomebodyMakesThis
AskReddit
books
webdesign
wallpapers
asoiaf
doctorwho
facebook
ideasfortheadmins
webcomics
ImaginaryLandscapes
smallbusiness
Lovecraft
soccer
AdviceAnimals
tipofmytongue
Documentaries
TheHobbit
occupywallstreet
blog
gonewild
woodworking
education
metacirclejerk
listentothis
AnythingGoesPics
FifthWorldPics
Accounting
Cascadia
secretsanta
teenagers
crappymusic
nfl
MMA
hookah
tattoos
triathlon
Seahawks
canada
washingtondc
hiphopheads
gonenatural
trailerparkboys
P90X
Drugs
philadelphia
BurningMan
GreenBayPackers
newjersey
Metalcore
ladybonersgw
gonewildcurvy
mashups
RedditLaqueristas
piercing
TrollXChromosomes
Luther
makinghiphop
C25K
Metal
BlackMetal
electricdaisycarnival
yoga
trance
stonerrock
abletonlive

Limits of Modularity

- Cannot discern topic, only communities of nearby nodes
- Subreddits of the same category far apart on the graph will be in different classes
- Subreddits of different categories close by on the graph will be in the same class
- Subreddits can be devoted to multiple topics



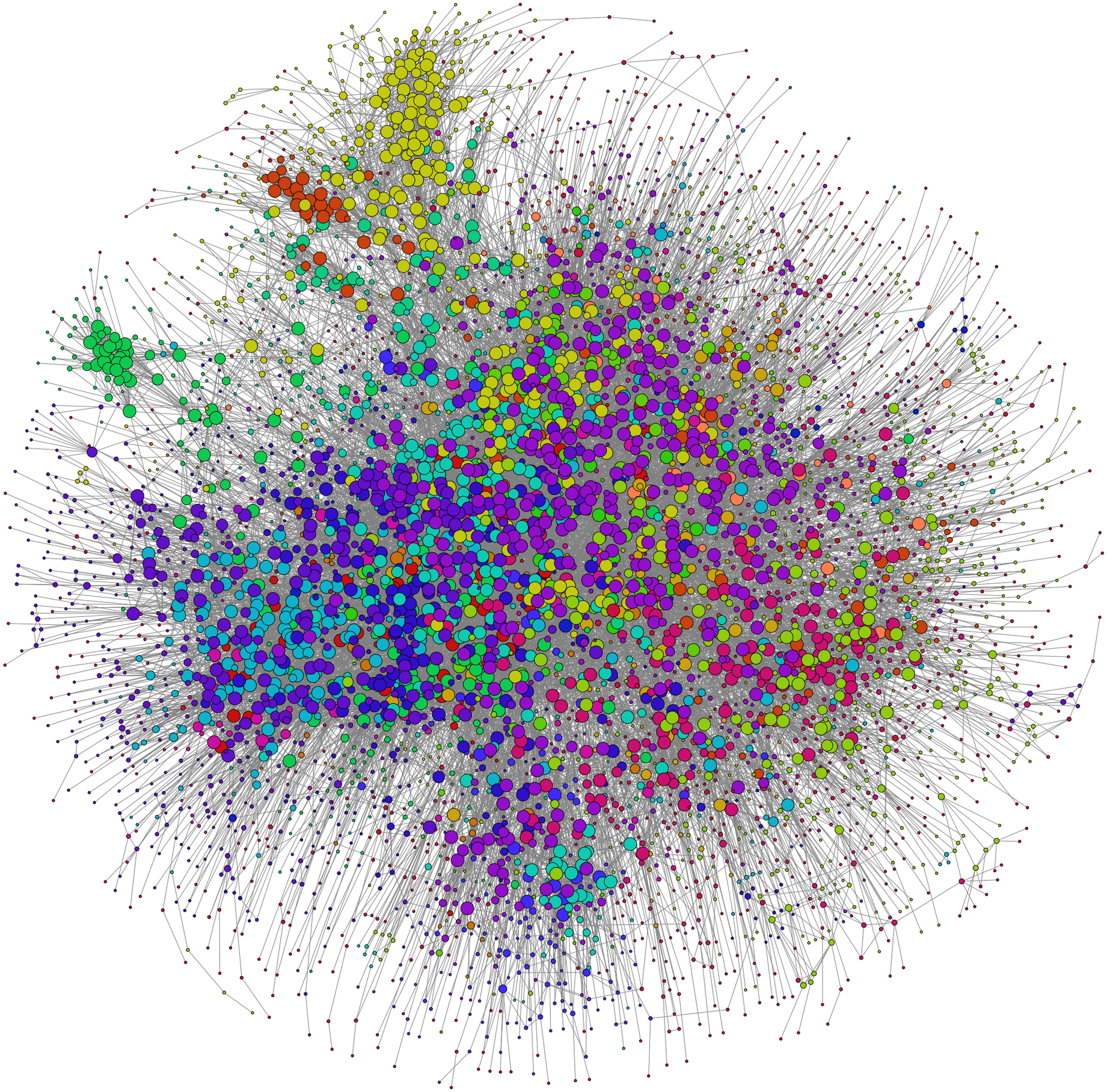


Algorithm Overview

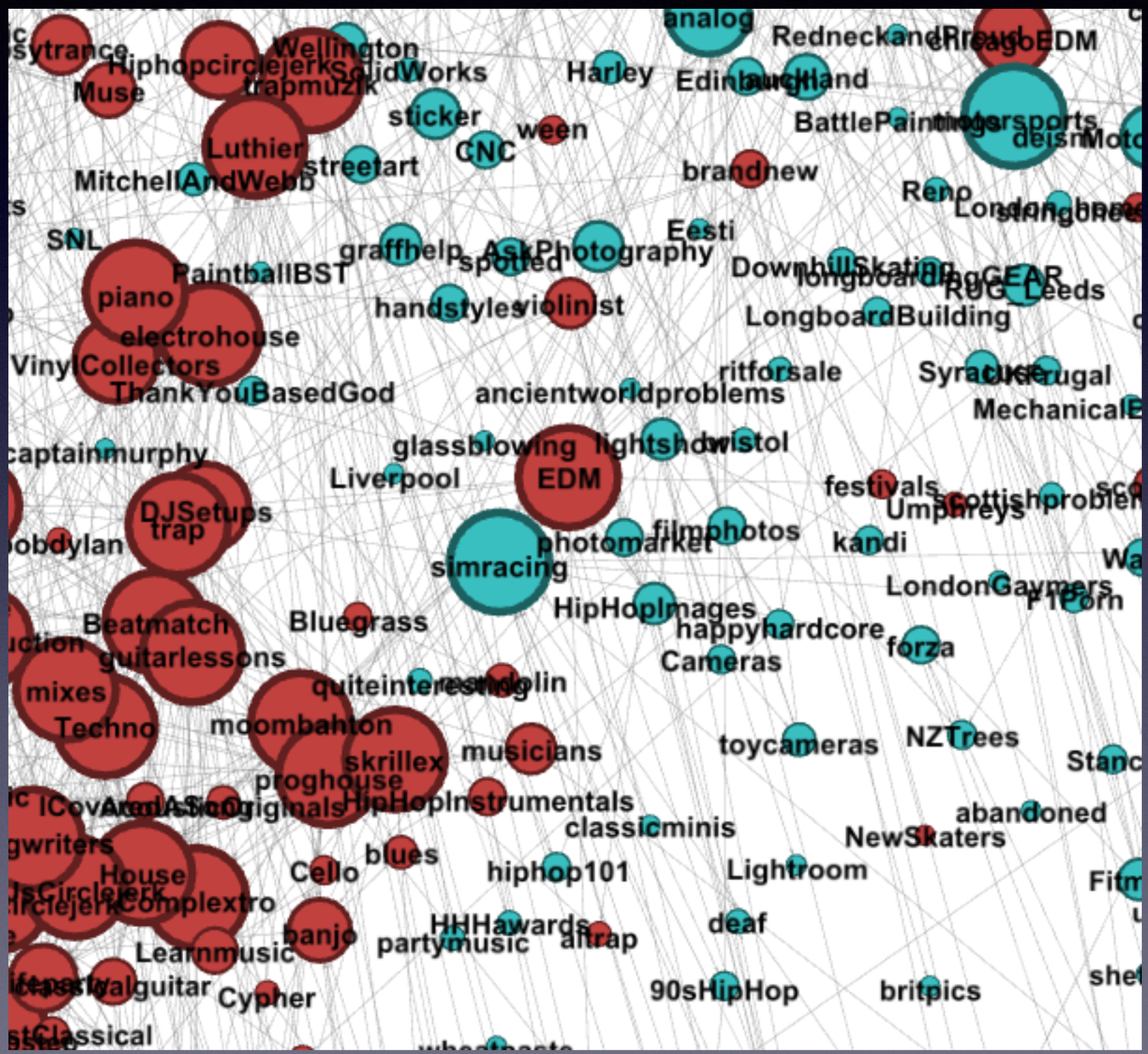
- Create list of topics (22)
  - music, meta, tech, politics, comedy, etc.
- Have the set of nodes be a |V|x22 array where each node has a degree of belonging to each category
- Manually create a seed set of categorized nodes (403)
- Use information diffusion process for each topic to “spread” category throughout the graph

Results

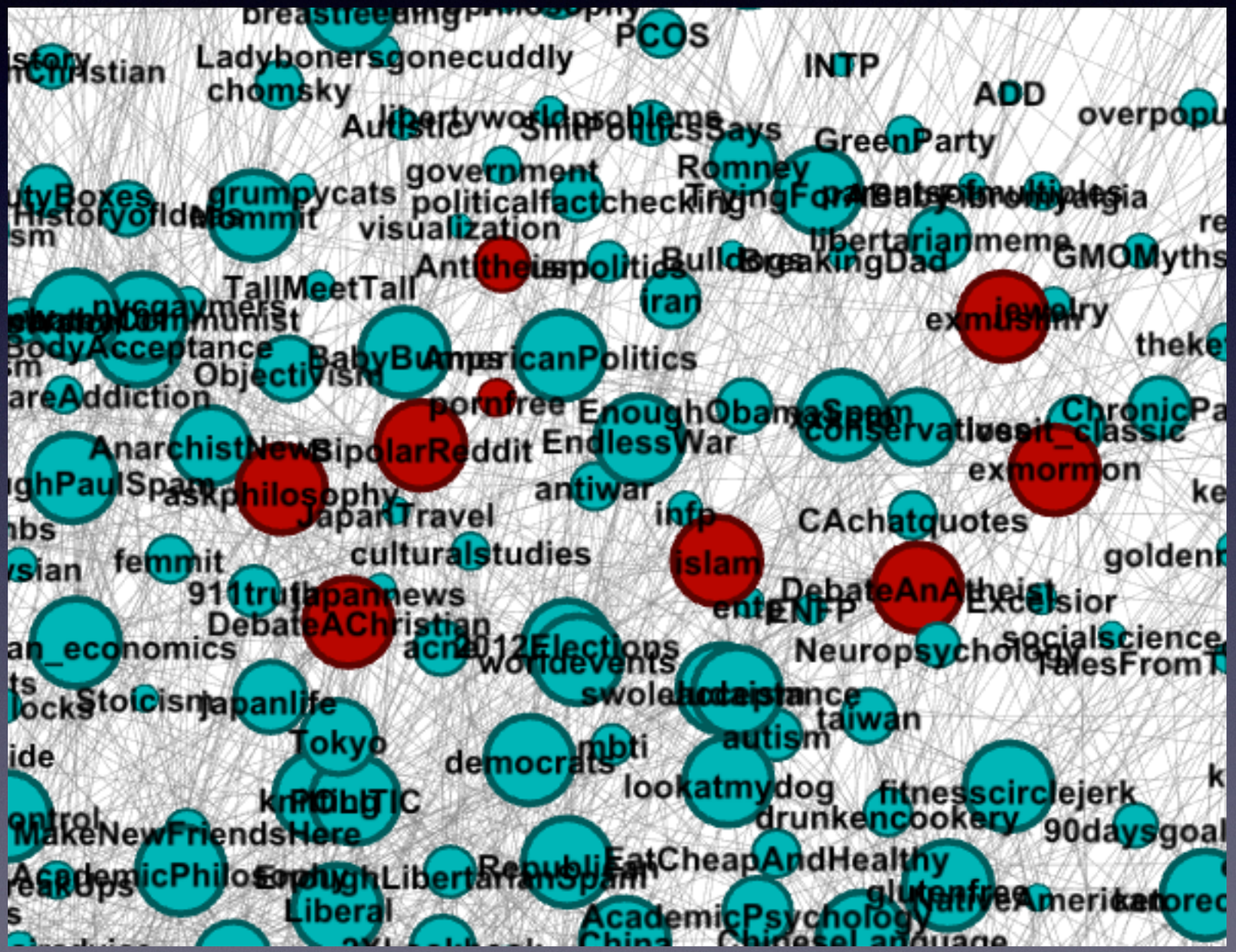
- Graph of each node's "primary category" shows a more nuanced, though more jumbled community structure than modularity
- Displaying each category individually gives better results
- Some categories diffused better than others



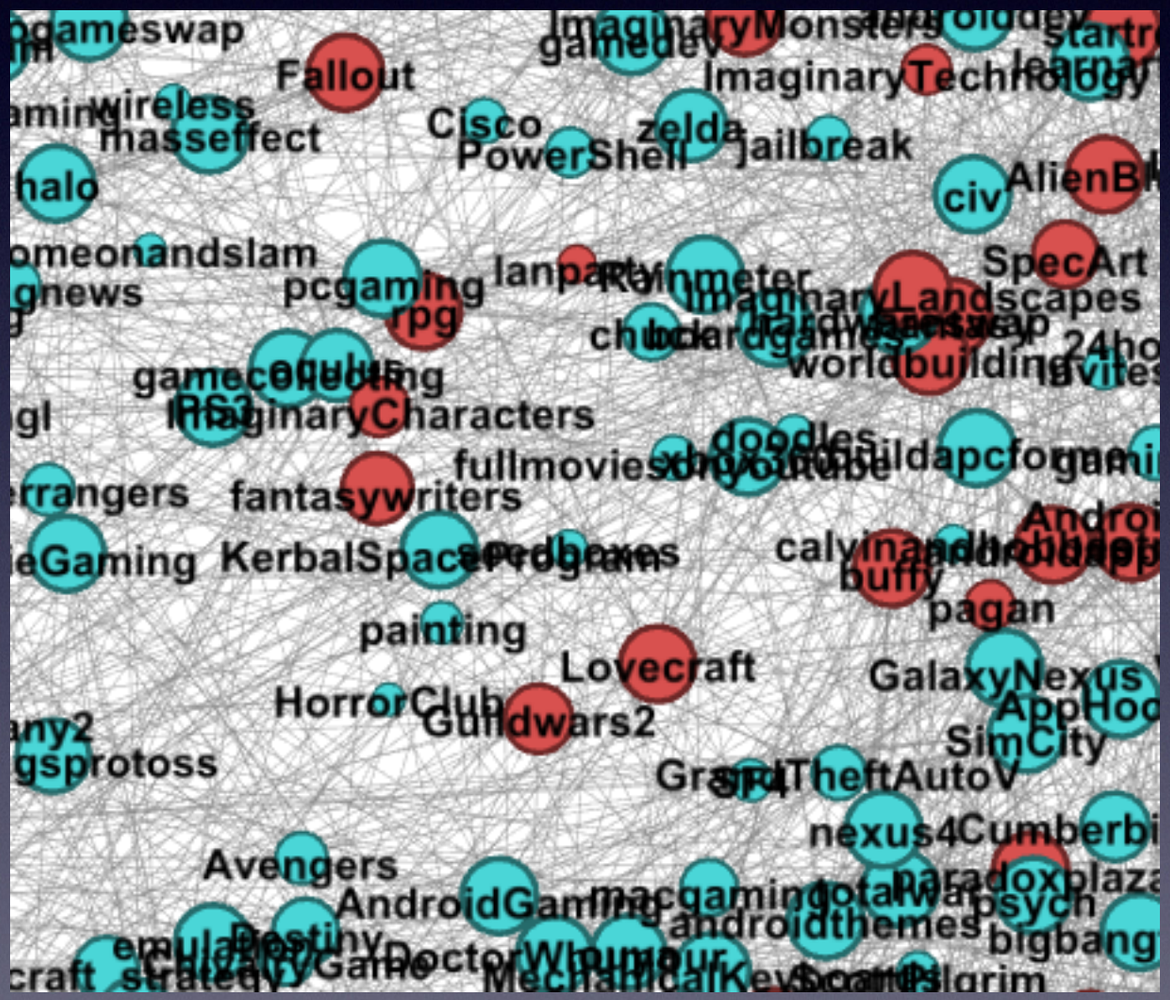
Music


analog
RednecksandRoads
chicagoEDM
Wellington
Hiphopcirclejerks
SolidWorks
Harley
Edinburgh
Auckland
Muse
trapmuzik
sticker
ween
BattlePaintings
CNC
Luthier
streetart
brandnew
MitchellAndWebb
Reno
SNL
Eesti
graffhelp
AskPhotography
spotted
DownhillSkating
PaintballBST
longboardingGEAR
RUG
Leeds
piano
handstyles
violinist
LongboardBuilding
electrohouse
VinylCollectors
ritforsale
Syracuse
ThankYouBasedGod
ancientworldproblems
MechanicalE
captainmurphy
glassblowing
lightshow
bristol
Liverpool
EDM
festivals
Umphreys
DJSetups
trap
filmphotos
photomarket
kandi
bobdylan
simracing
LondonGaymers
HipHopImages
Bluegrass
Beatmatch
happyhardcore
forza
guitarlessons
Cameras
mixes
quiteinteresting
Techno
moombahton
toycameras
NZTrees
skrillex
musicians
proghouse
HipHopInstrumentals
abandoned
classicminis
NewSkaters
blues
House
Cello
hiphop101
Lightroom
Complextro
deaf
banjo
partymusic
altrap
Learnmusic
Cypher
90sHipHop
britpics
Classical

Religion


PCOS
Ladybonersgonecuddly
chomsky
INTP
ADD
Autistic
GreenParty
government
Romney
grumpycats
politicalfactchecking
visualization
libertarianmeme
GMOMyths
Antitheism
uspolitics
TallMeetTall
iran
jewelry
BodyAcceptance
AmericanPolitics
Objectivism
pornfree
EnoughObamaSpam
conservatives
exmormon
AnarchistNews
BipolarReddit
EndlessWar
askphilosophy
antiwar
JapanTravel
infp
CAchatquotes
culturalstudies
islam
femmit
911truth
japannews
DebateAChristian
DebateAnAtheist
ENFP
Excelsior
2012Elections
worldevents
Neuropsychology
socialscience
Stoicism
japanlife
taiwan
Tokyo
autism
democrats
mbti
lookatmydog
fitnesscirclejerk
drunkencookery
MakeNewFriendsHere
EatCheapAndHealthy
LibertarianSpain
Liberal
AcademicPsychology
glutenfree

Books/Literature



Improvements & Future Work

- Better seed set
  - more balanced, more correct, more data
  - pick central nodes
- Better categories
- Optimize algorithm parameters
- Implement dictionary/semantic similarity algorithm?
- Take posts into account
- Improve diffusion algorithm

Credits

- Olson, Randal S. & Neal, Zachary P. “Navigating the massive world of reddit: Using backbone networks to map user interests in social media.” December 12, 2013