Reddit network analysis

Ema Pajić, Nikola Aleksić, Predrag Obradović, Marko Mišić, Bill Power, Zoran Obradović

Introduction

- **Reddit** is a social news aggregation, web content rating, and discussion website.
- Posts are organized by subject into user-created boards called **communities** or **subreddits**.
- Some stats:
  - 50M+ daily active users
  - 100K+ active communities
  - 50B+ monthly views

Motivation

- Many possible research directions using Reddit data
- Analyzing online conflicts – Community Interaction and Conflicts on the Web
- Hate speech analysis - Dreaddit: A Reddit Dataset for Stress Analysis in Social Media
- Impact of real life events (Covid, elections, finance/crypto)
- Analyzing how people use the website - Generalists and Specialists: Using Community Embeddings to Quantify Activity Diversity in Online Platforms

- **Prior work either doesn't use user overlap, or uses basic variants**

Pushshift

- **Pushshift Reddit API** collects Reddit data every day and provides easy access to everyone.
- Accessing data via **crawler API** or **large monthly dumps**

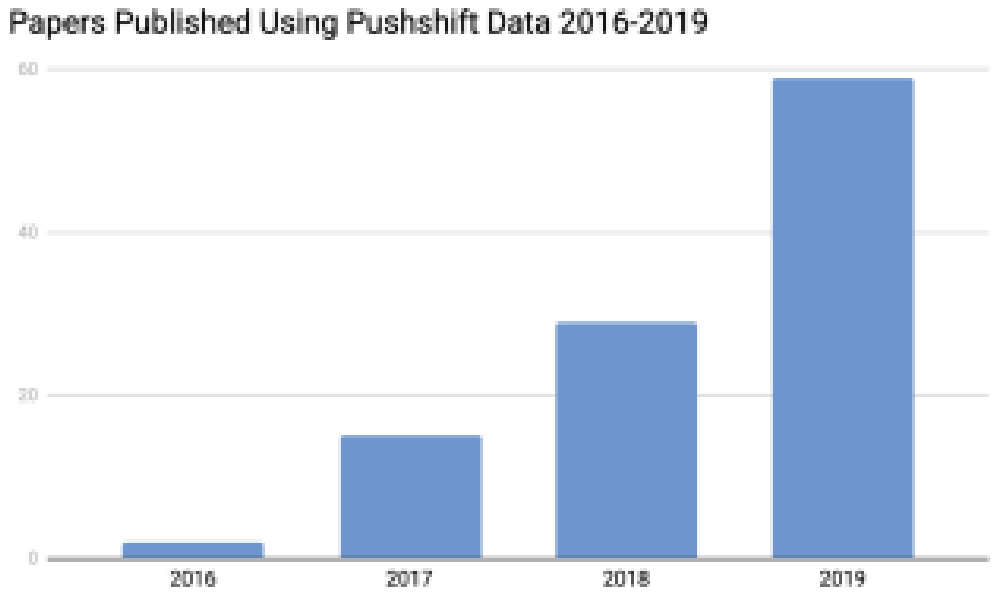
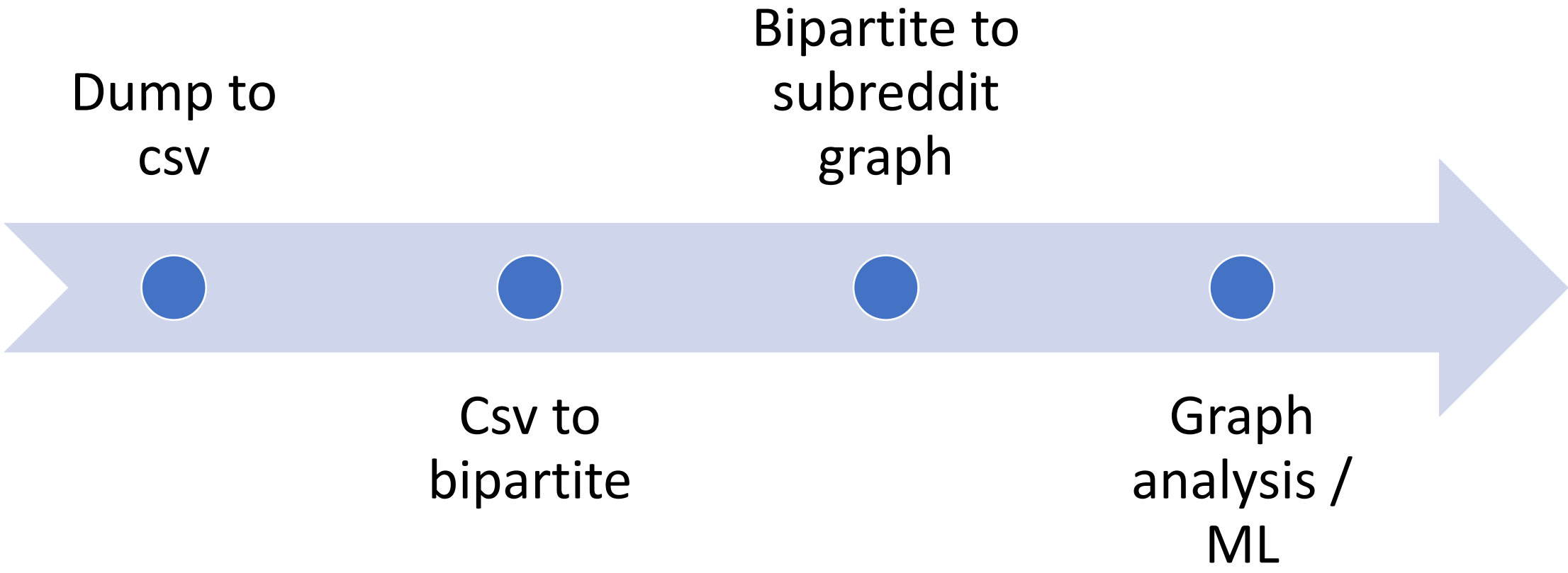


Figure from **The Pushshift Reddit Dataset Paper**

Datasets

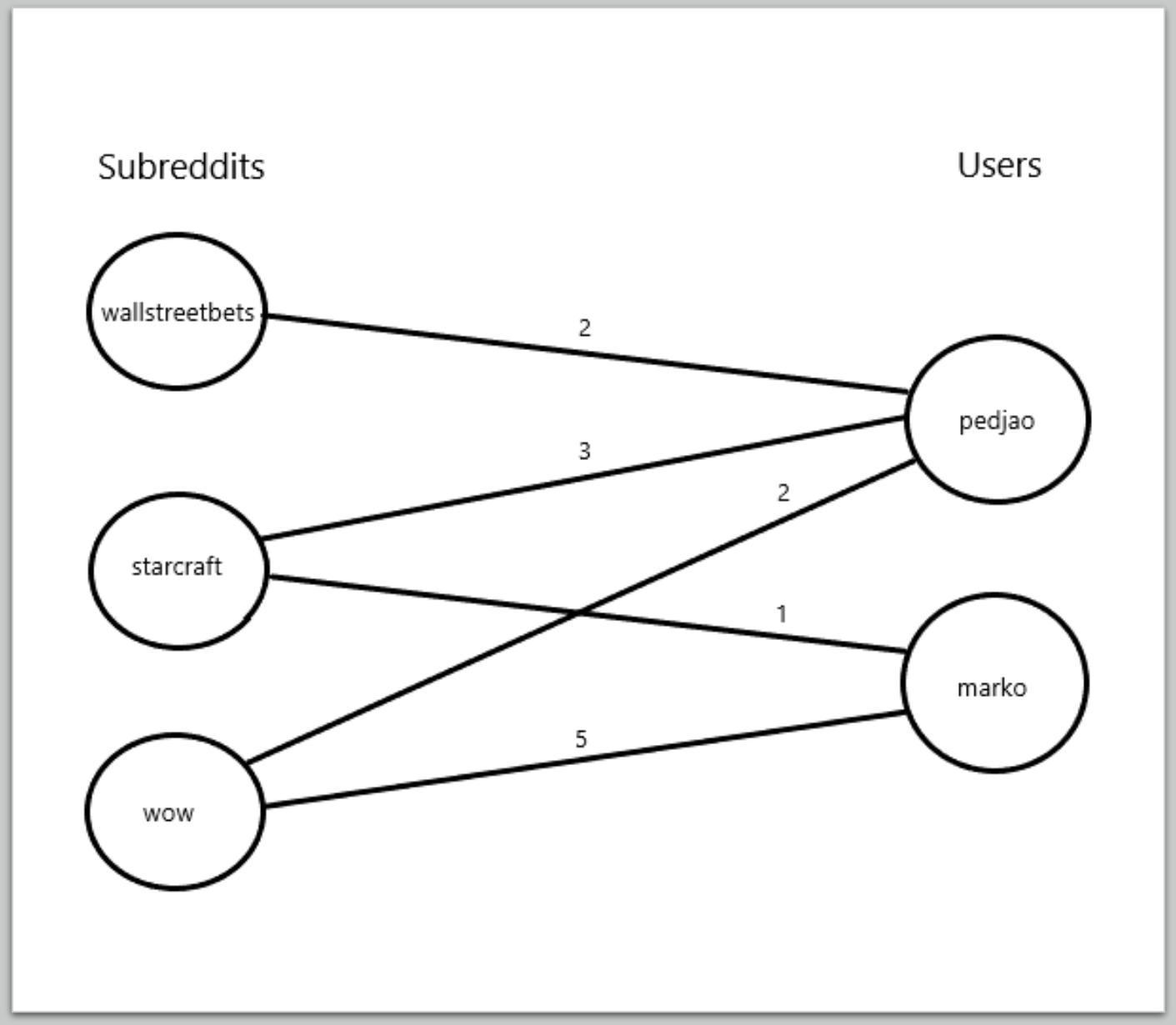
- Crawling data
  - Possible through Pushshift API
  - pmaw, psaw
  - Not reliable
- Pushshift dumps
  - December 2005 – June 2021
  - One dump contains all comments and posts posted in a month
  - 2021 – around 20GB compressed, 200GB uncompressed for each month

Pipeline

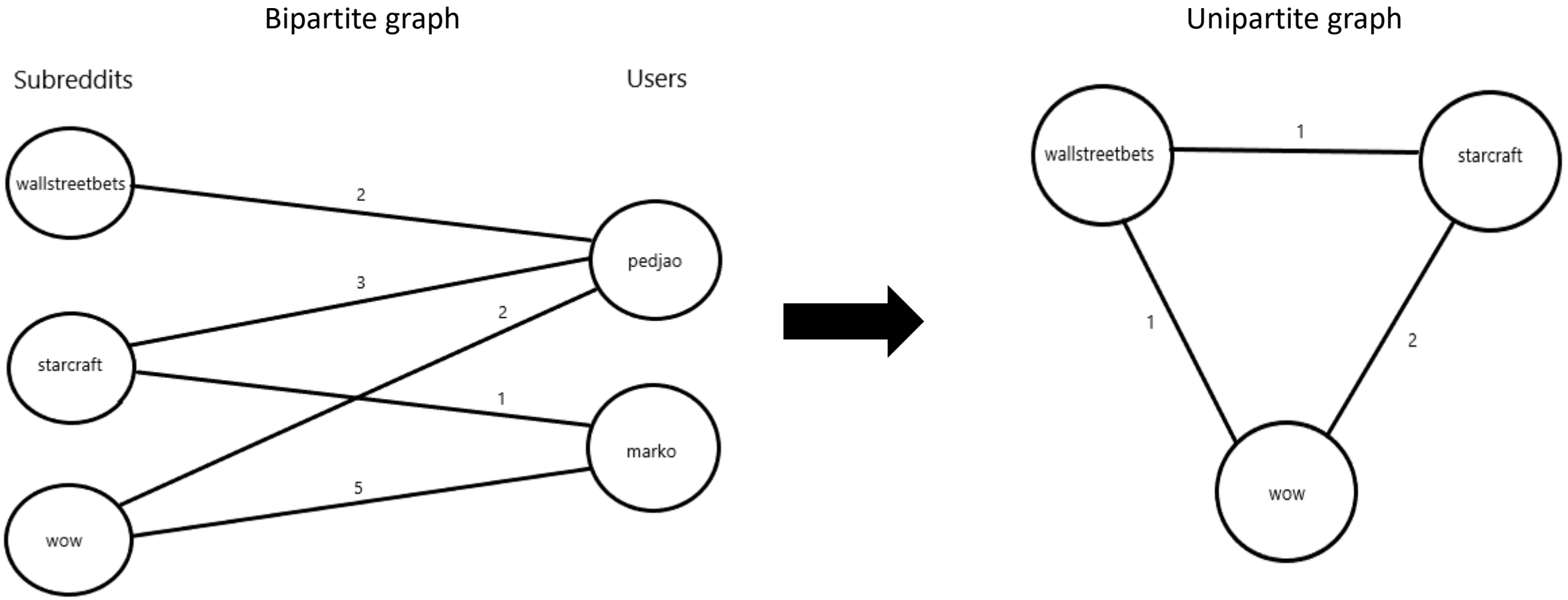

Dump to csv
Csv to bipartite
Bipartite to subreddit graph
Graph analysis / ML

Modelling Reddit

- Bipartite graph – Weight between **Subreddit** and **User** equal to number of comments **User** posted on that **Subreddit**
- Many different ways to transform into Unipartite graph



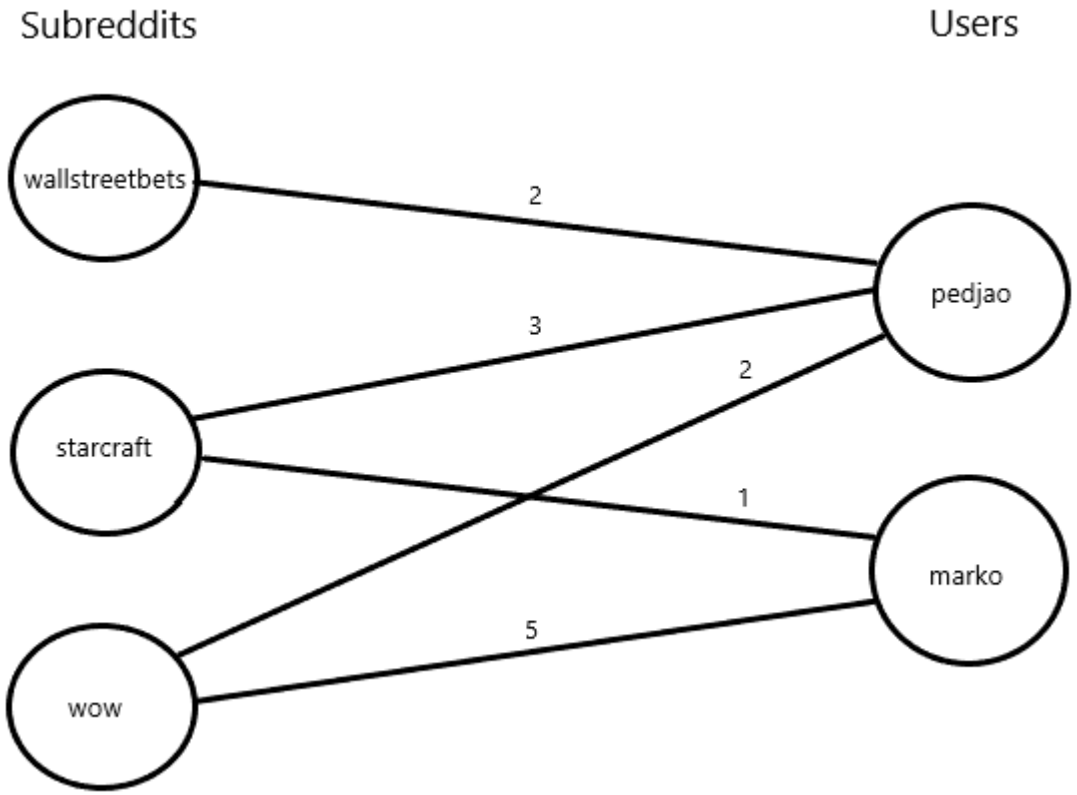
Count overlapping users



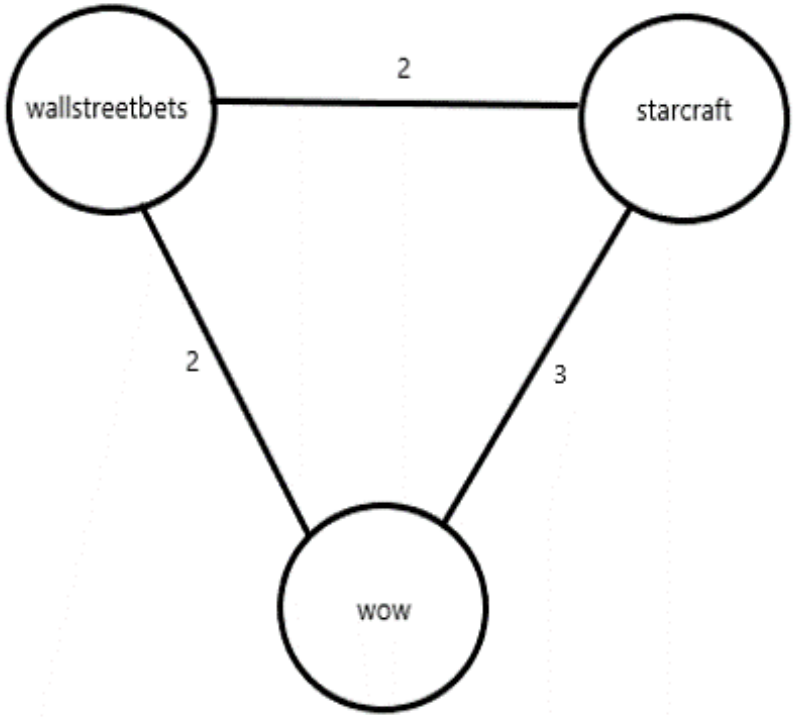
Problem: Two subreddits might have one user posting 1 time on both, while another user posted 50 and 60 times on each, respectively. We don't want to count these the same!!!

Sum minimum between user weights

Bipartite graph



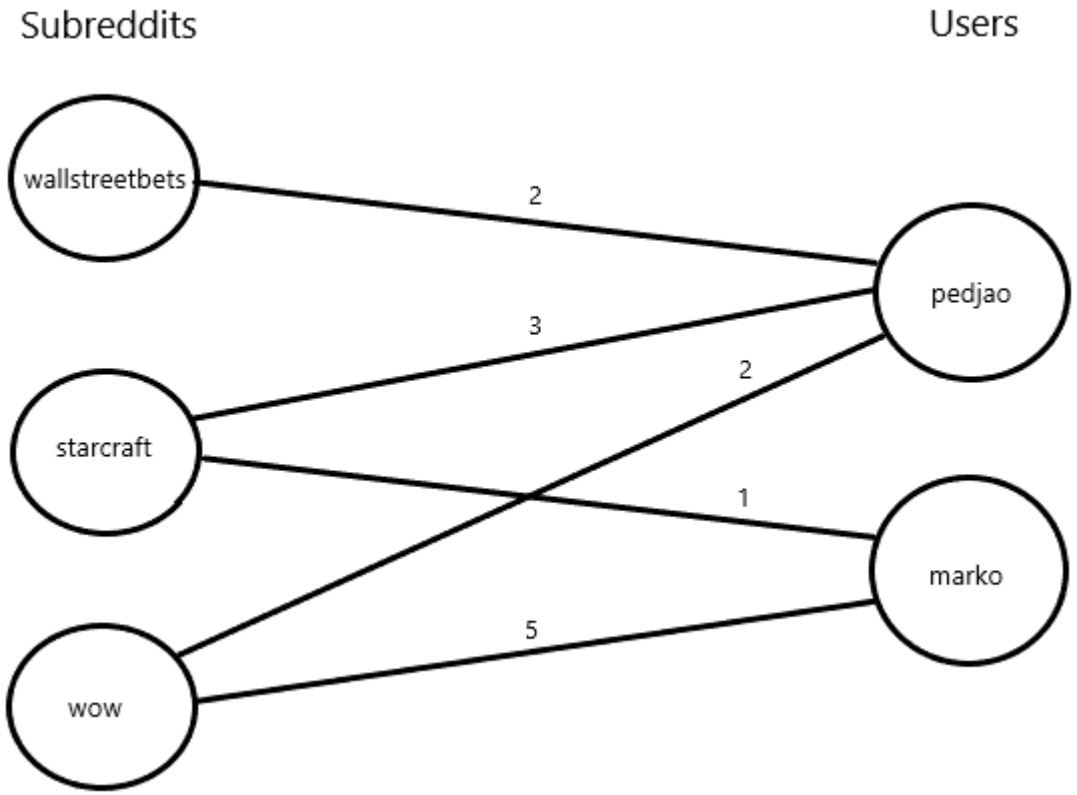
Unipartite graph



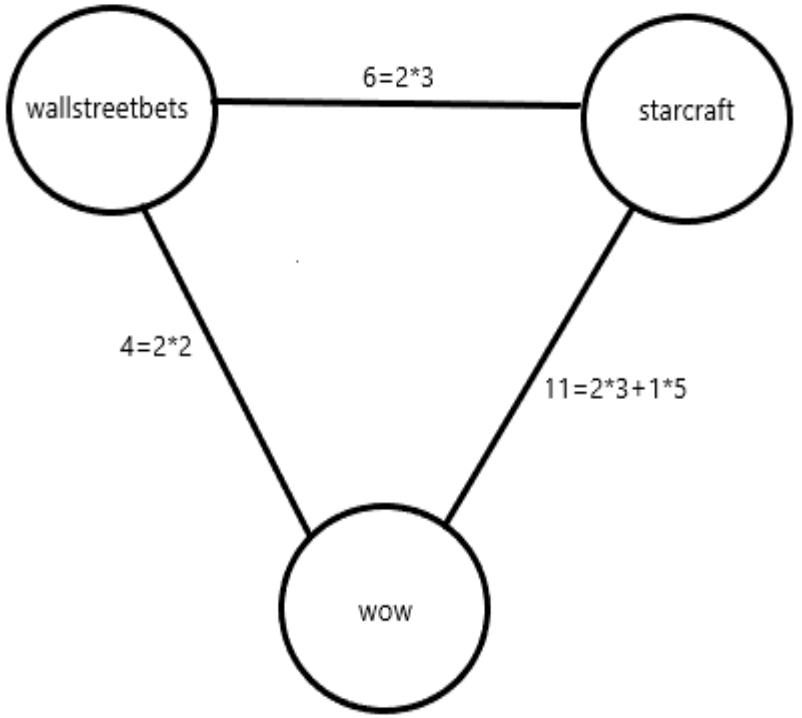
Problem: Strong connection between **marko** and **wow** is disregarded. Also, **pedjao** doesn't visit **starcraft** and **wow** the same but the weight to **wallstreetbets** is the same.

Sum product between user weights

Bipartite graph



Unipartite graph

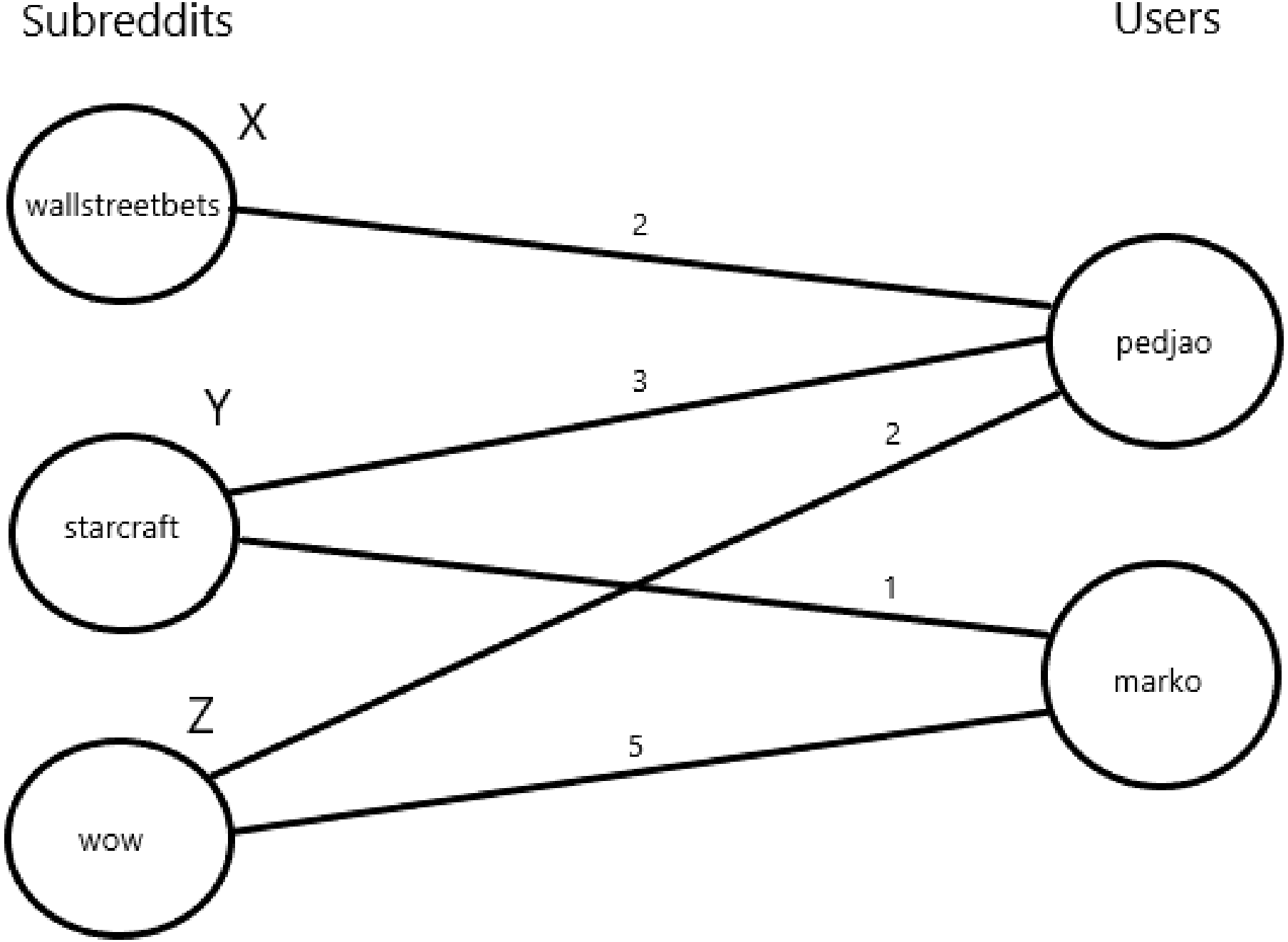


Problem: Weights getting too big. What does the weight of 6 mean, is it big or small, does it correspond to a strong connection?

Two pass aggregation[1]

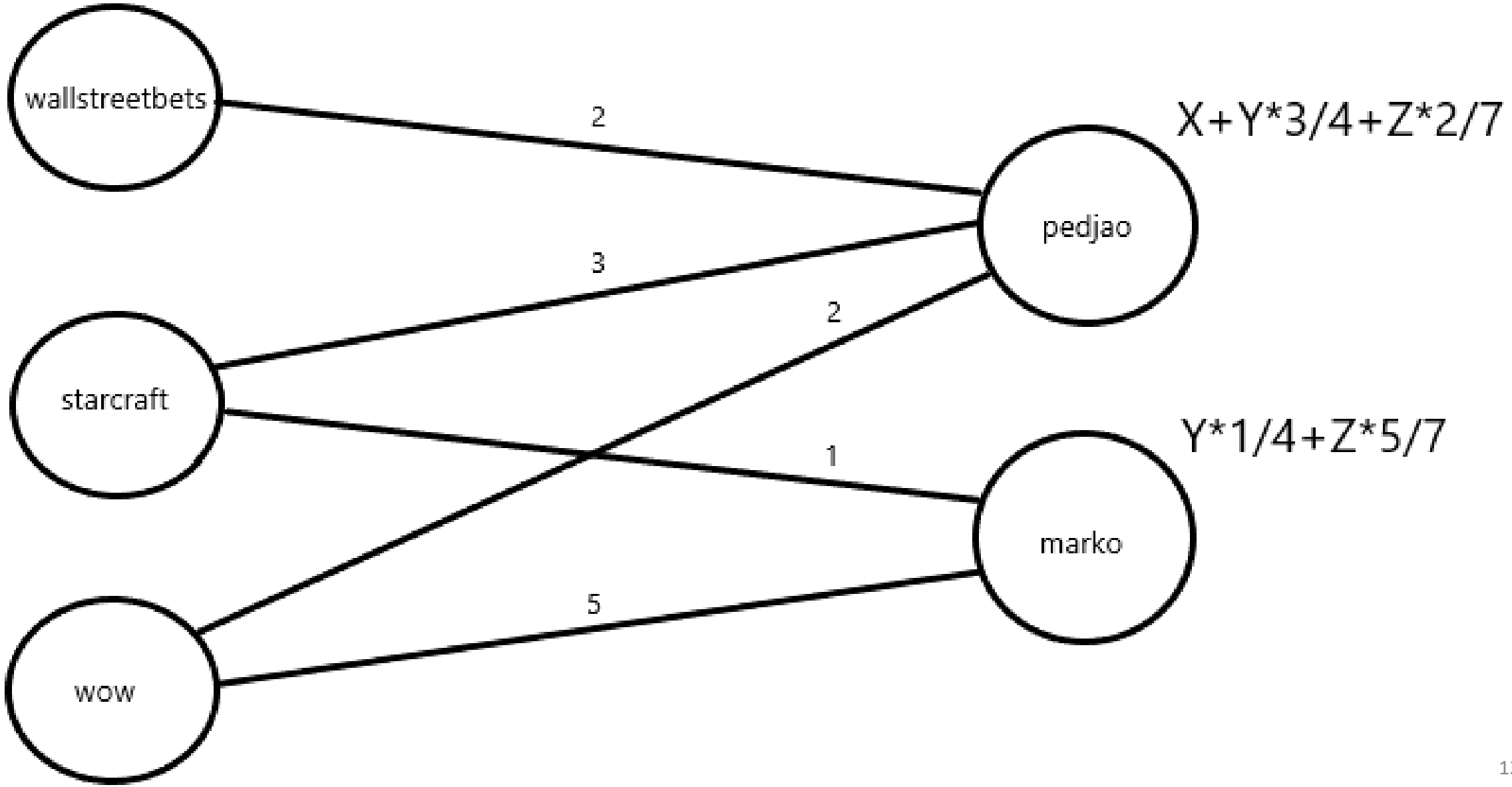
1. Assign 1 unit of **resource** to a subreddit.
2. Redistribute resource to connected users, splitting it proportionally to edge weights.
3. Every user does the same, redistributes the weight to subreddits, proportionally to edge weights.
4. Amount of resource that started in subreddit1 and ended in subreddit2 we will be the edge weight!

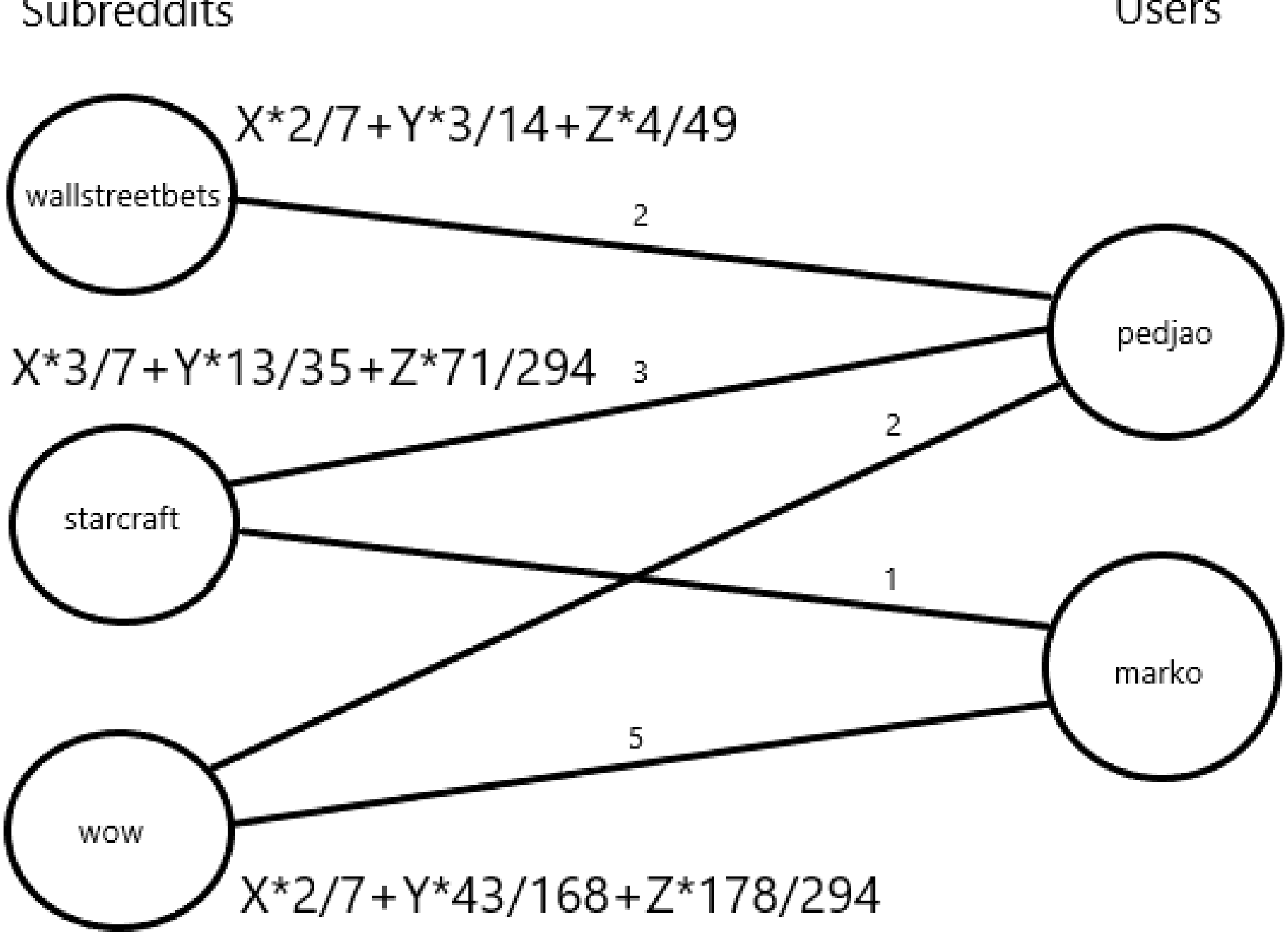
1.https://www.researchgate.net/publication/5851936_Bipartite_network_projection_and_personal_recommendation


Subreddits
Users
X
wallstreetbets
Y
starcraft
Z
wow
pedjao
marko
2
3
2
1
5

Subreddits

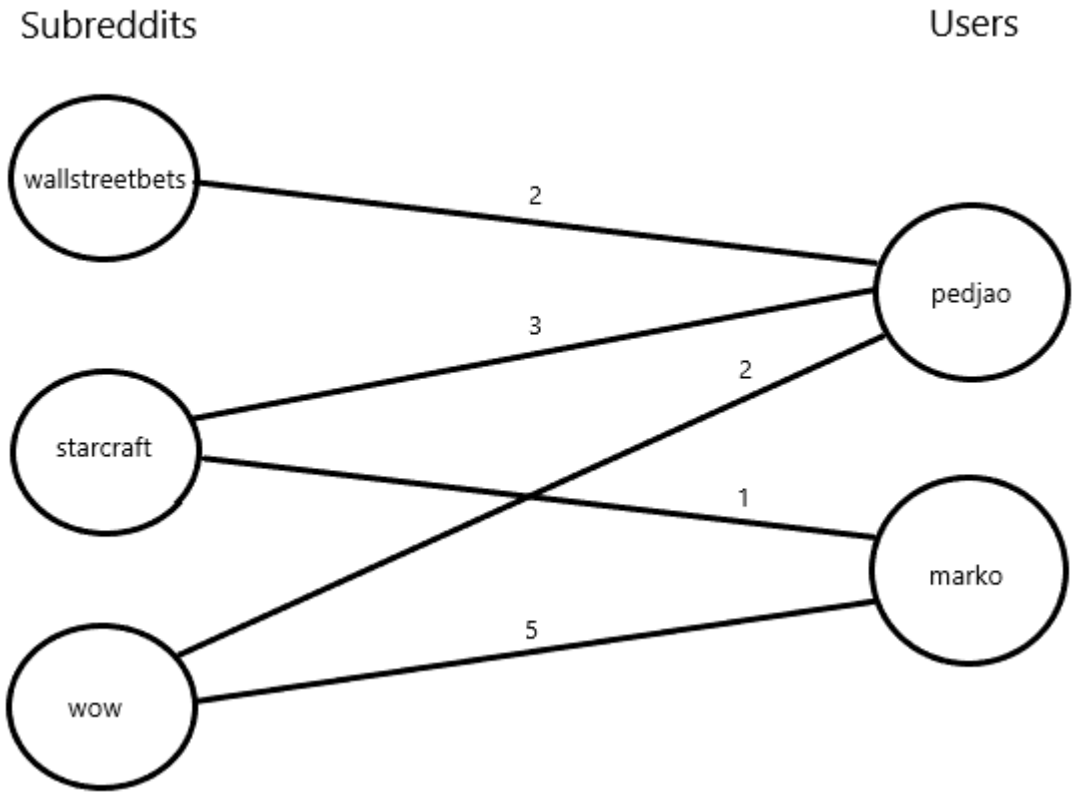
Users



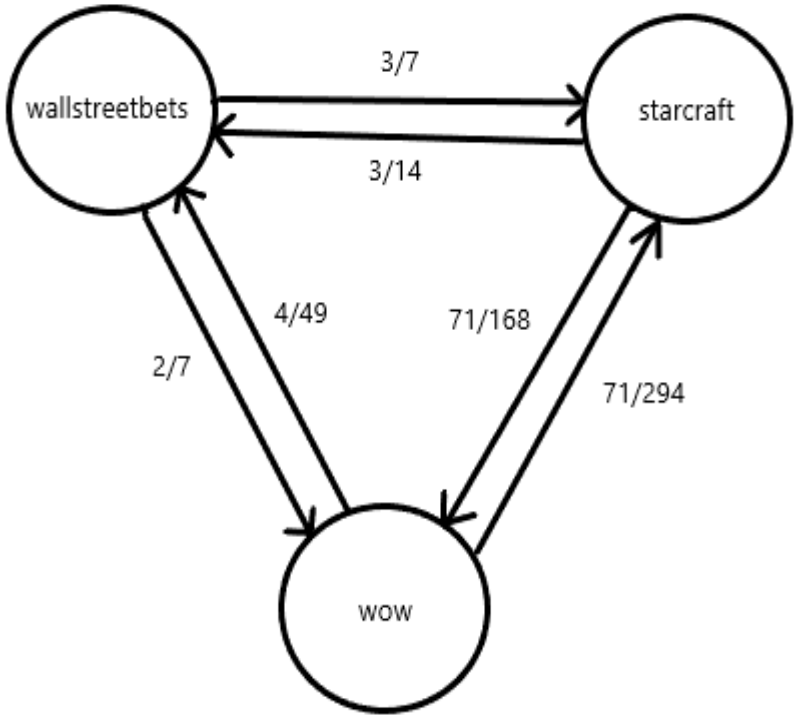

Subreddits
Users
wallstreetbets
X*2/7+Y*3/14+Z*4/49
2
pedjao
X*3/7+Y*13/35+Z*71/294
3
2
starcraft
1
marko
5
wow
X*2/7+Y*43/168+Z*178/294

- Two pass aggregation

Bipartite graph



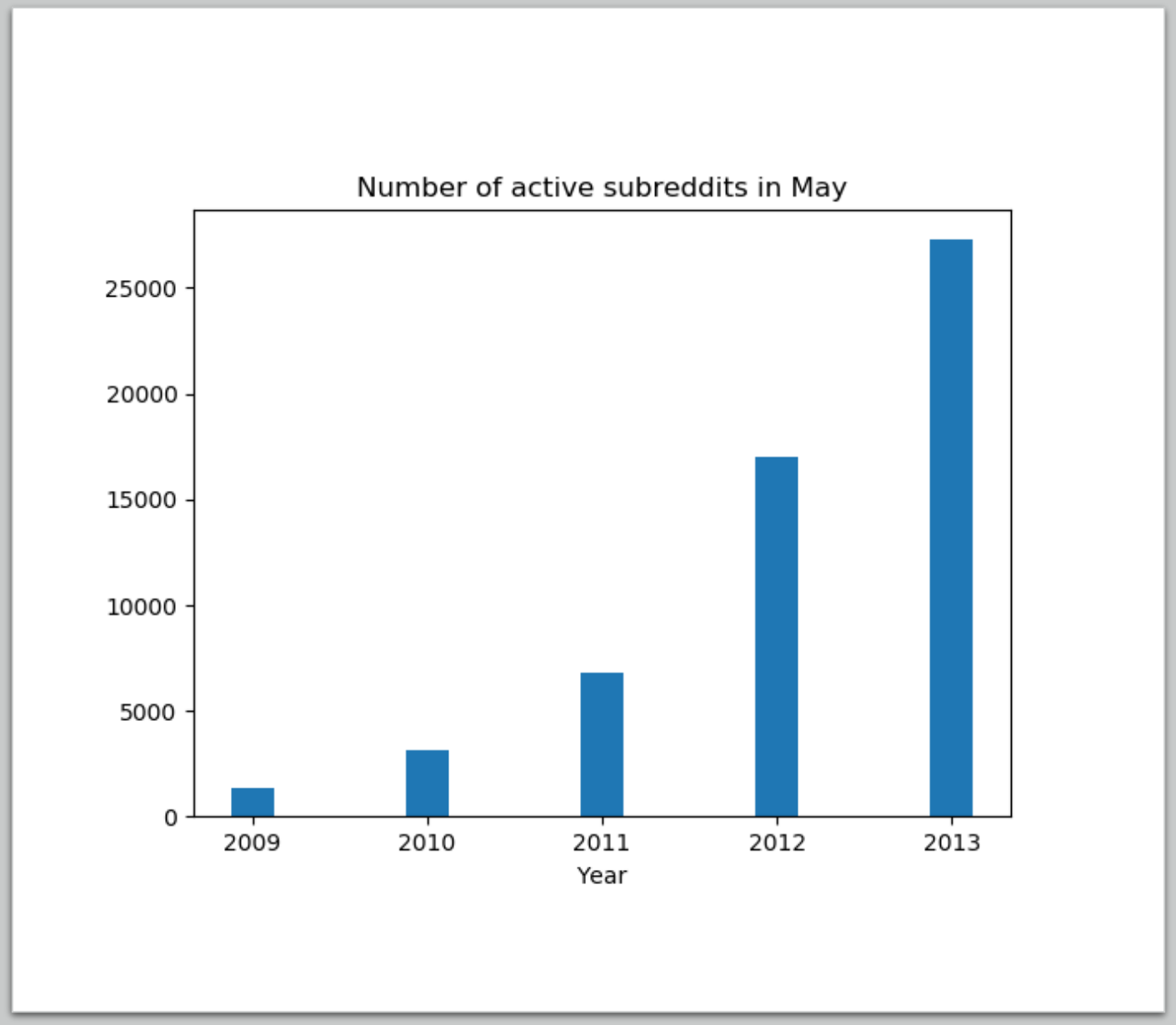
Unipartite graph

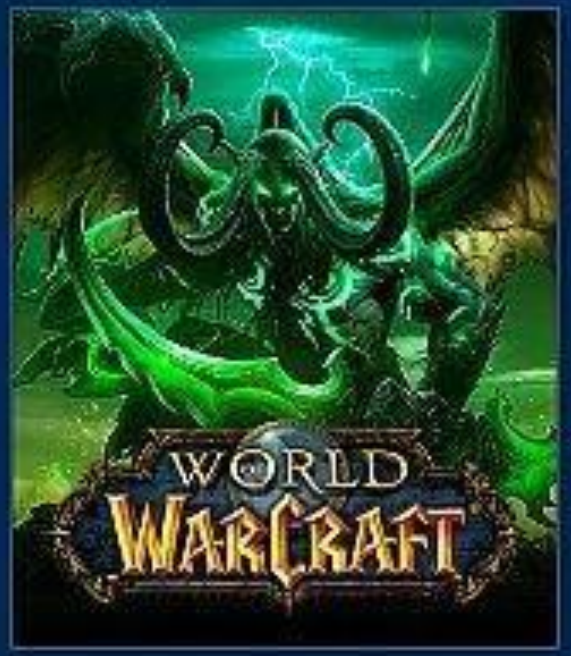


Awesome properties: Similar to using product as earlier, but normalizes by both user and subreddit sizes. **Weights are between 0 and 1!!!**

Applying on real data

- Reddit data is huge
- Even for one month in early Reddit years, $10^6 - 10^9$ edges.
- To perform analysis in NetworkX, limit on one topic




WORLD
WARCRAFT


OVERWATCH


HEARTHSTONE
HEROES OF WARCRAFT


DIABLO


HEROES
OF THE STORM

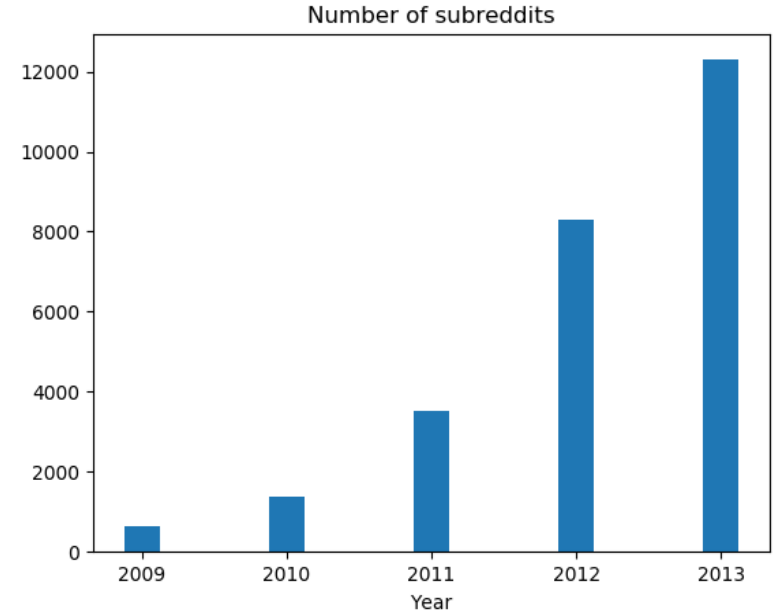

STARCRAFT
II

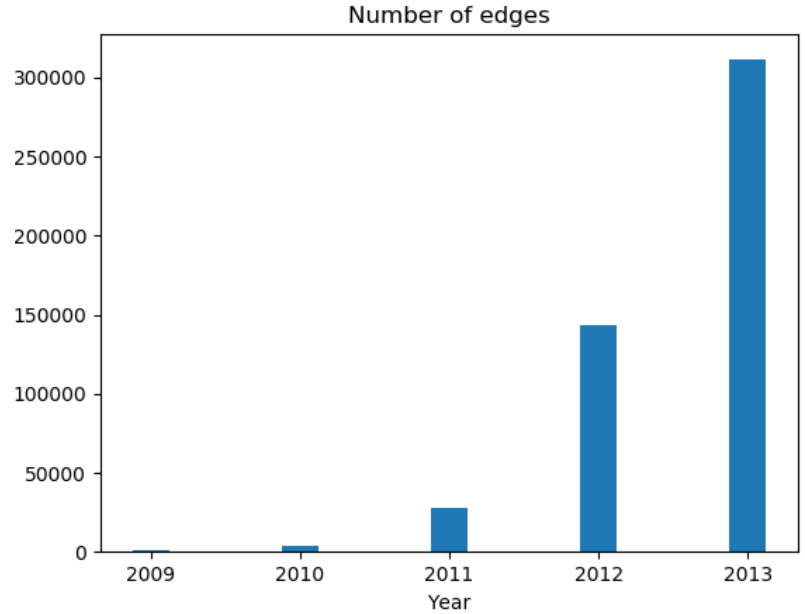

BLIZZARD
CLASSICS

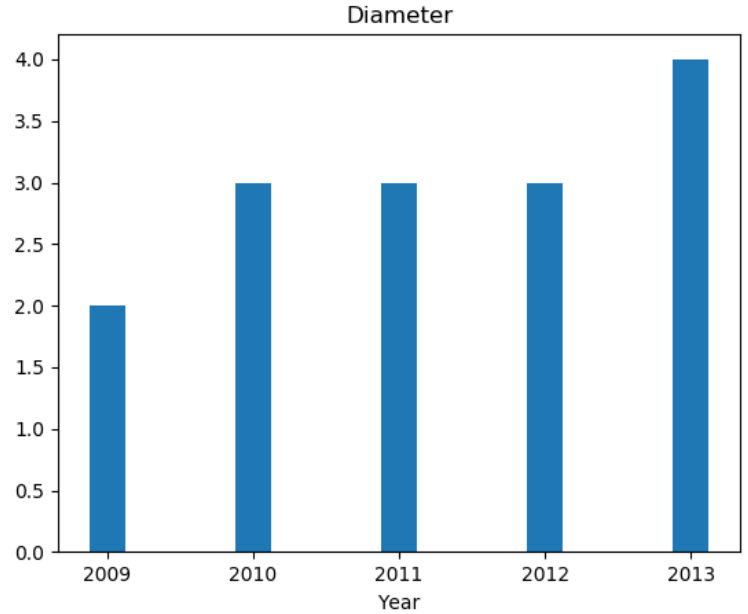

BATTLE.NET

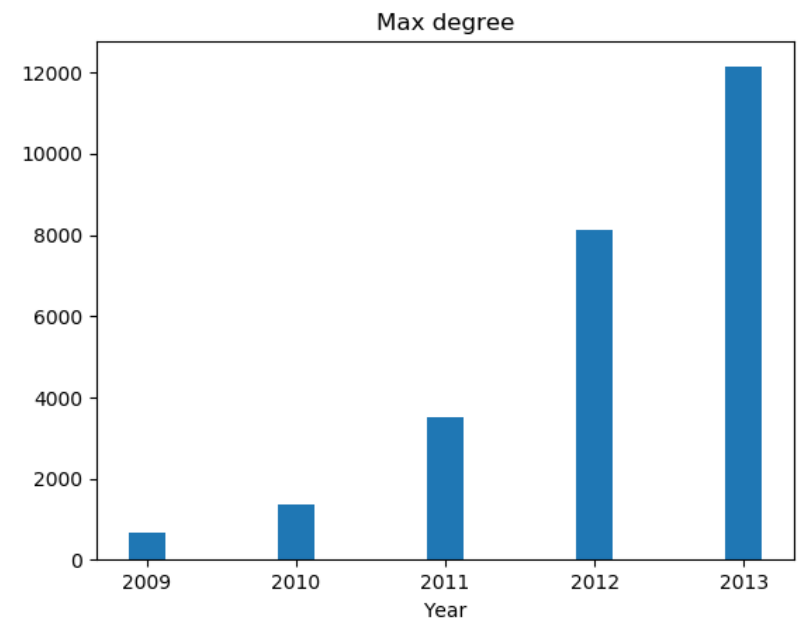
Blizzard subreddit analysis

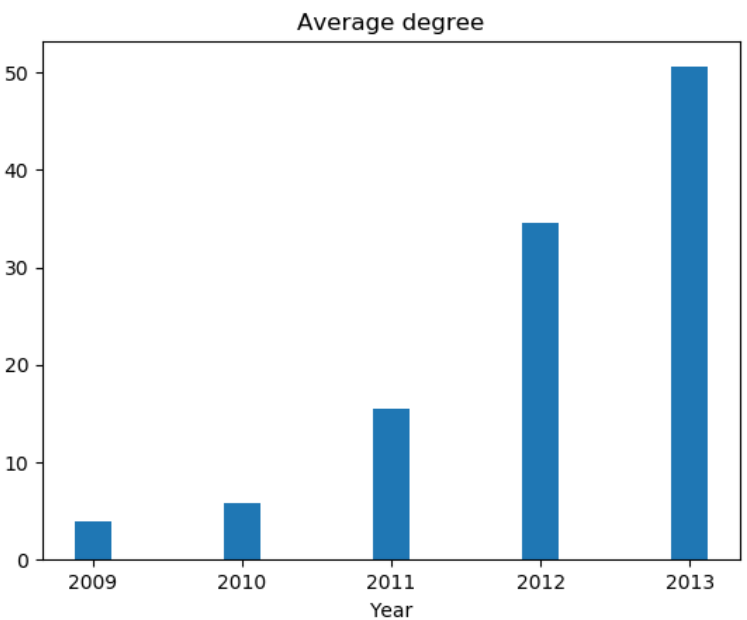
- **Blizzard –** one of the biggest video game companies
- Manually collected ~ 100 of Blizzard subreddits
- Computed edges between Blizzard subreddits, and Blizzard subreddits and other subreddits
- Added connected subreddits to the list
- Analyzing only May 2009-2013

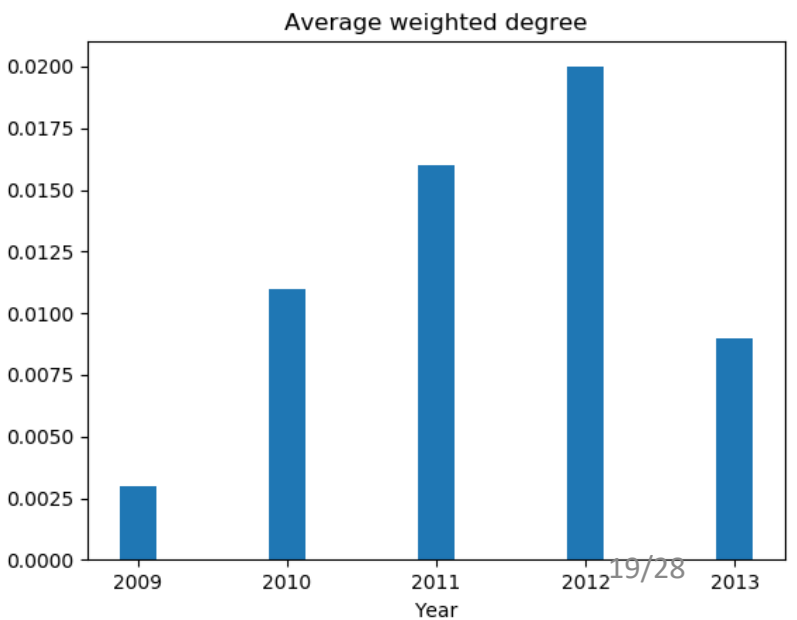

Number of subreddits
12000
10000
8000
6000
4000
2000
0
2009
2010
2011
2012
2013
Year


Number of edges
300000
250000
200000
150000
100000
50000
0
2009
2010
2011
2012
2013
Year


Diameter
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2009
2010
2011
2012
2013
Year


Max degree
12000
10000
8000
6000
4000
2000
0
2009
2010
2011
2012
2013
Year


Average degree
50
40
30
20
10
0
2009
2010
2011
2012
2013
Year


Average weighted degree
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
2009
2010
2011
2012
2013
Year

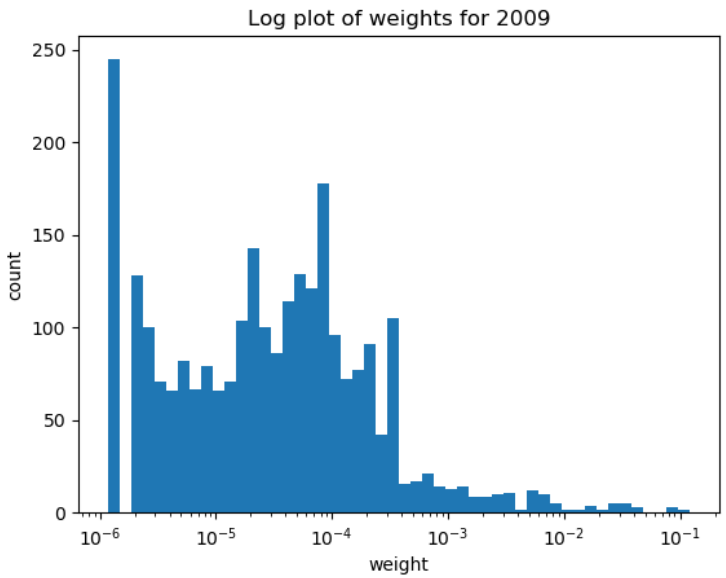
Extracting a backbone

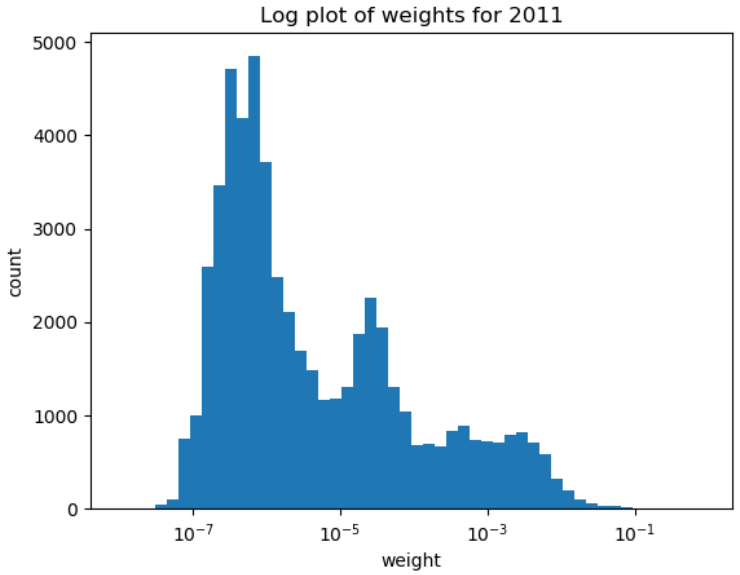
- After projecting a bipartite graph, we are left with many really weak edges
- **Keep strong edges, remove weak ones**
- Threshold approach: $\begin{cases} weight > T, keep\ the\ edge \\ weight < T, remove\ the\ edge \end{cases}$
- More advanced approaches: Look at a distribution of that edge weight for large number of randomly generated graphs, keep the edge if the weight is statistically significant.[1]

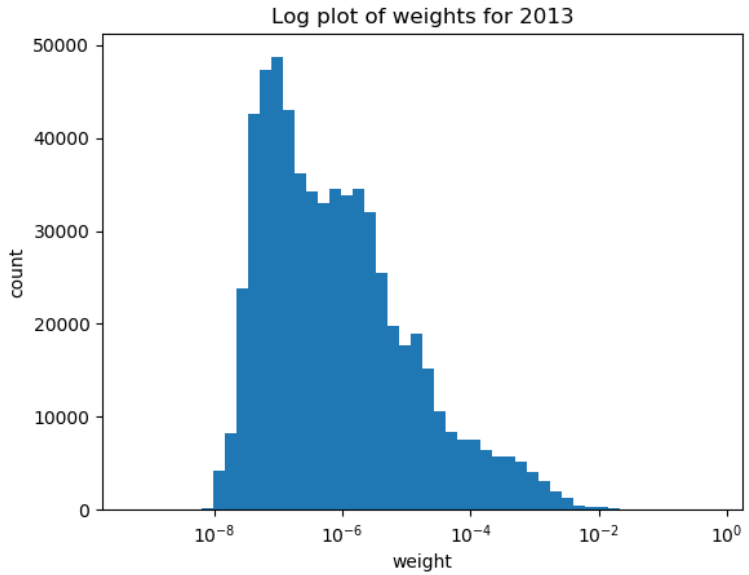
1.https://arxiv.org/abs/2105.03391

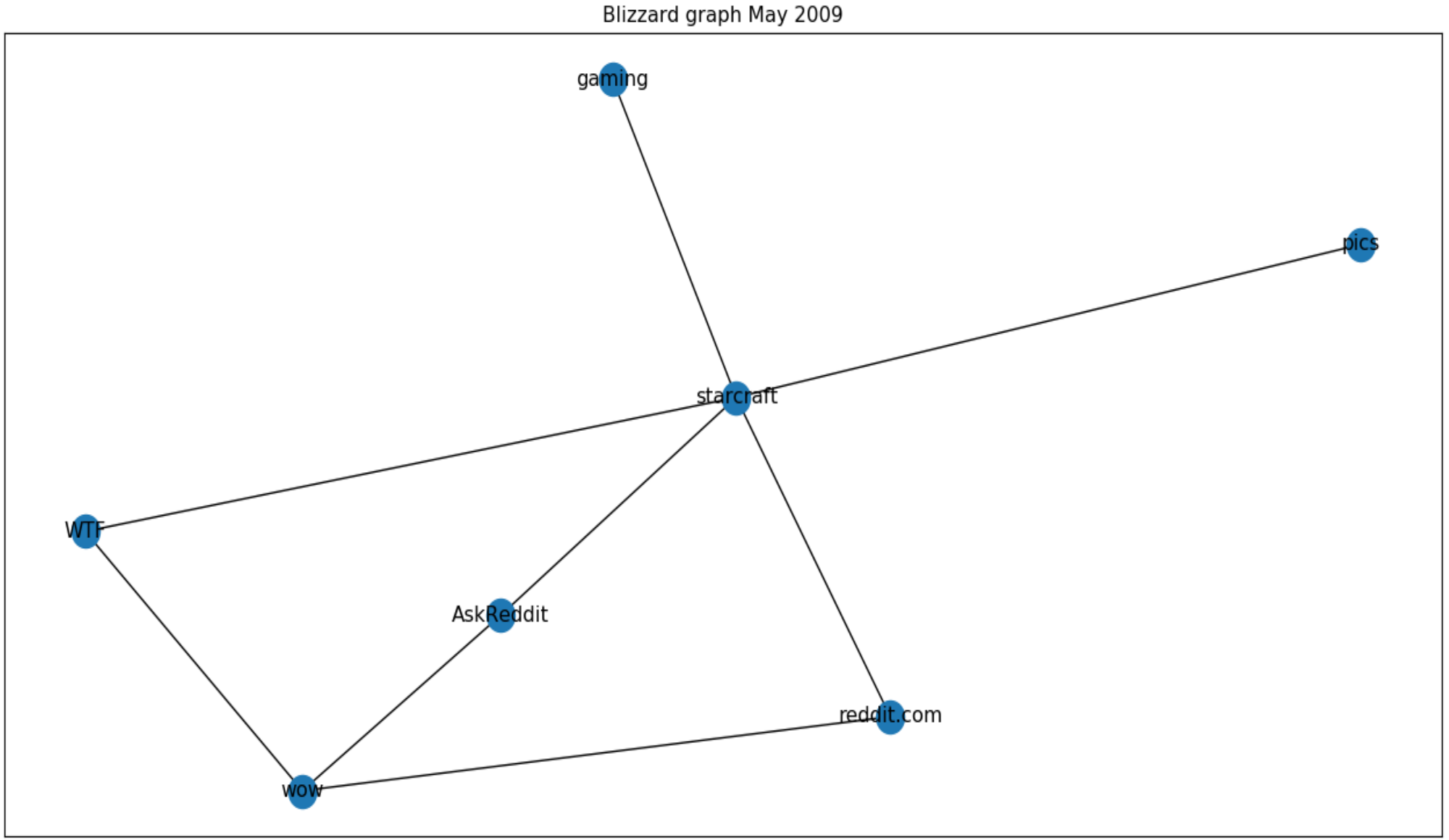
Choosing the threshold

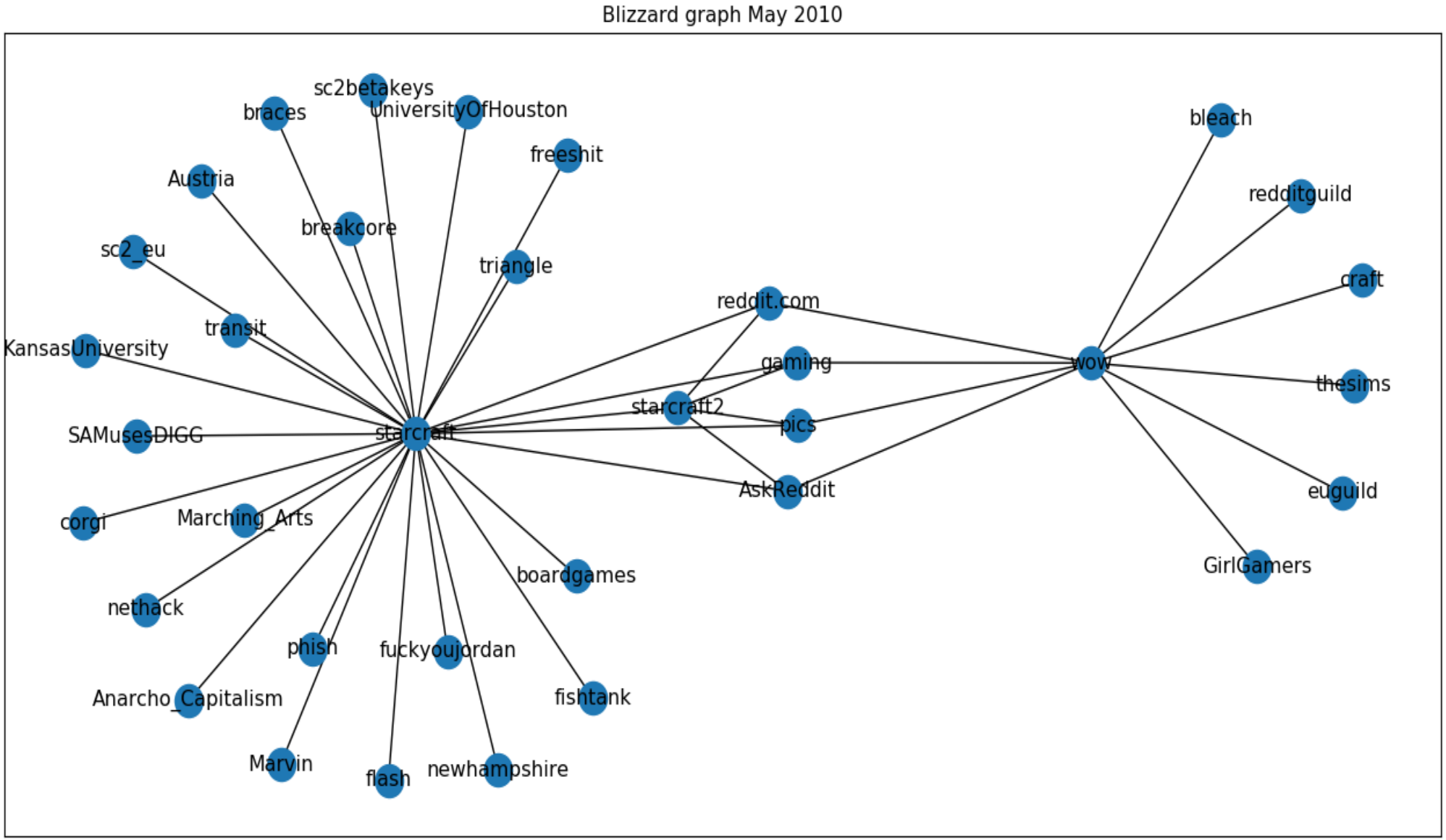
- As we want a small number of important edges, we need to choose a strict threshold that removes >99% edges
- We see that values between 10^{-2} and 10^{-1} satisfy that. We choose 0.04.

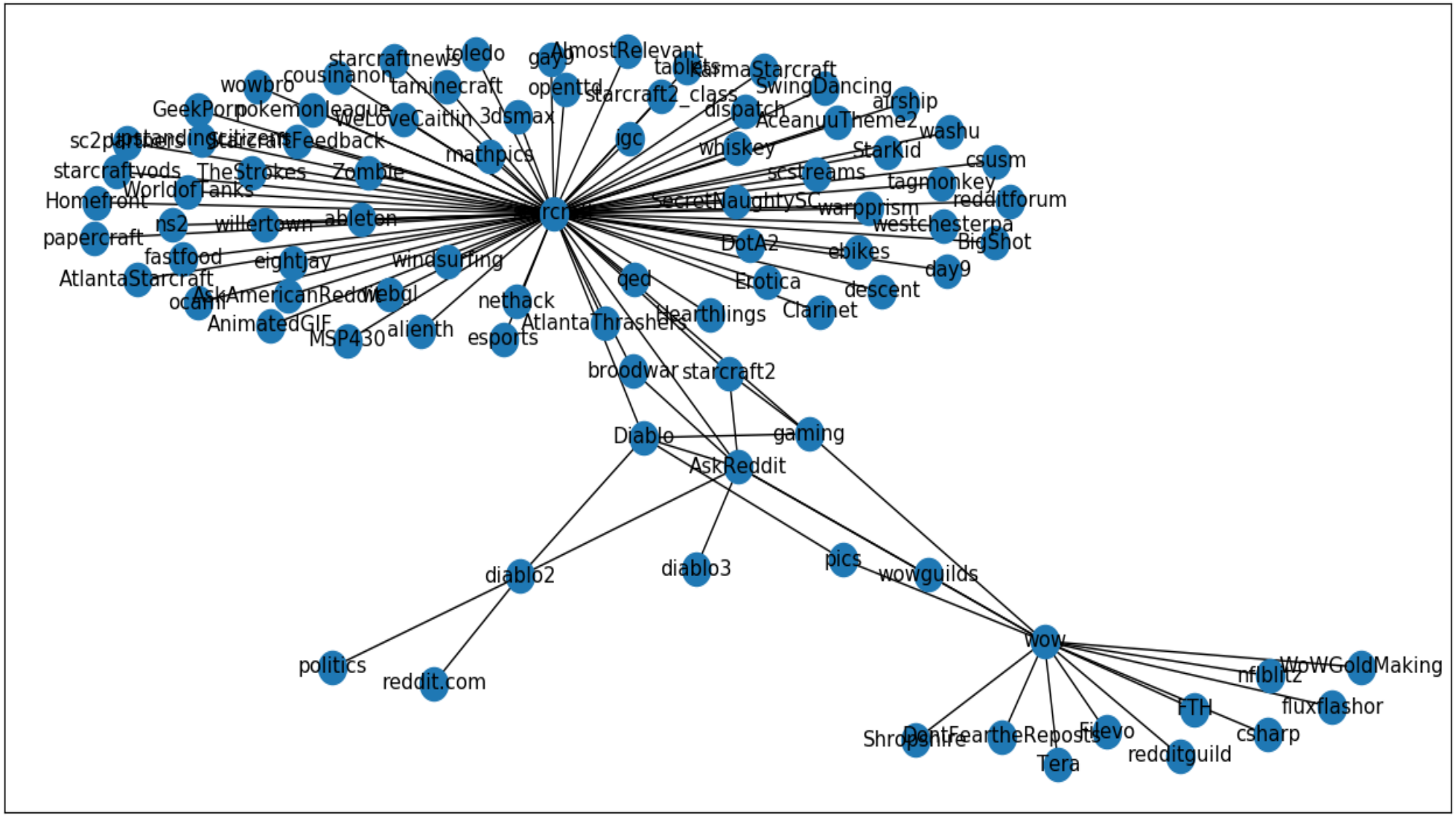


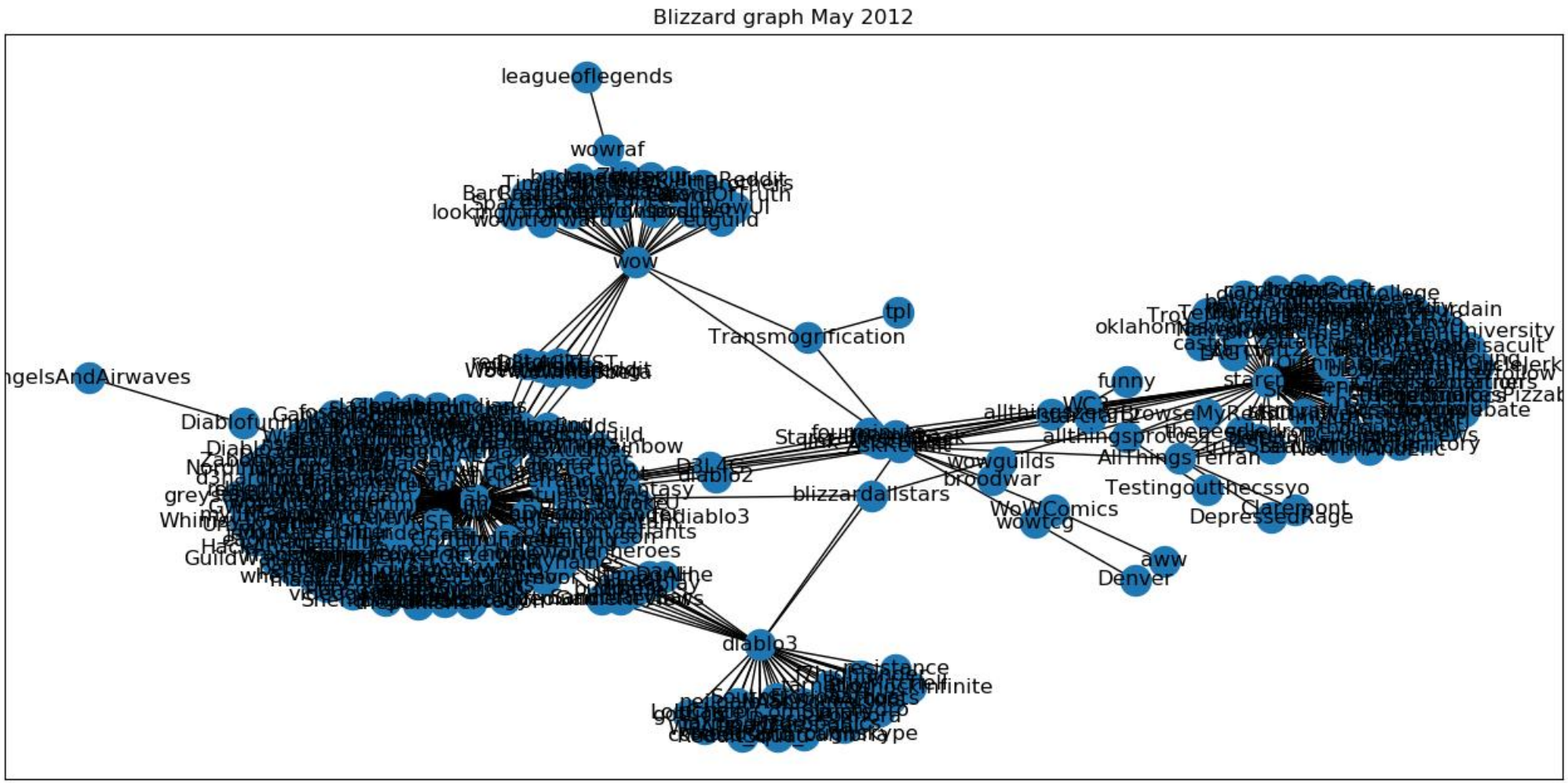


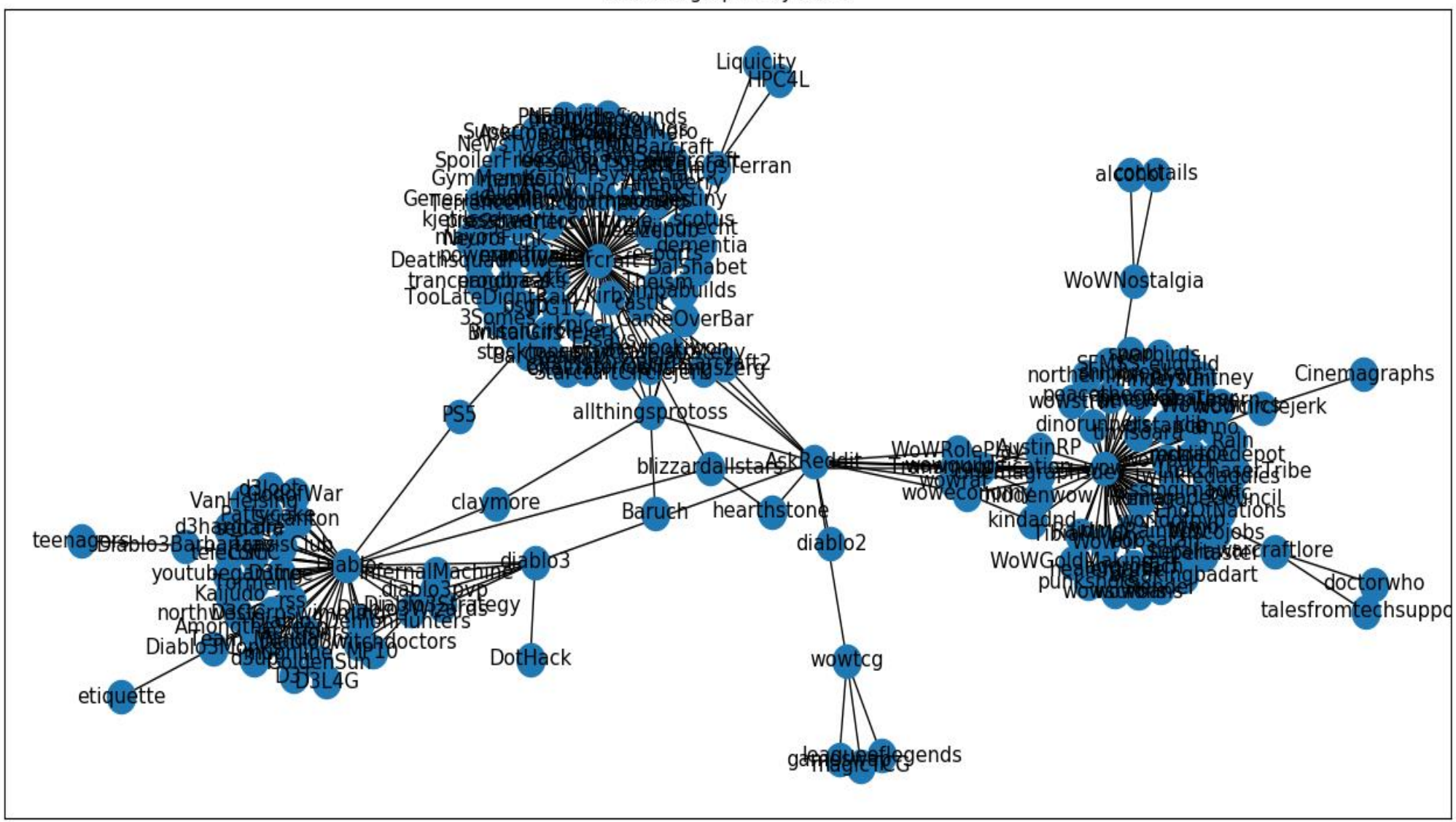



Blizzard graph May 2009
gaming
pics
starcraft
WTF
AskReddit
reddit.com
wow


Blizzard graph May 2010
sc2betakeys
braces
UniversityOfHouston
freeshit
Austria
breakcore
sc2_eu
triangle
transit
KansasUniversity
SAMusesDIGG
starcraft
corgi
Marching_Arts
nethack
boardgames
phish
fuckyoujordan
Anarcho_Capitalism
fishtank
Marvin
flash
newhampshire
reddit.com
gaming
starcraft2
pics
AskReddit
bleach
redditguild
craft
wow
thesims
euguild
GirlGamers


Blizzard graph May 2011
starcraftnews
toledo
gay
AlmostRelevant
cousinanon
tamingcraft
openttd
tablets
KarmaStarcraft
wowbro
starcraft2_class
SwingDancing
GeekPorn
pokemonleague
dispatch
airship
WeLoveCaitlin
3dsmax
AceanuuTheme2
washu
sc2pantsers
StarCraftFeedback
igc
whiskey
StarKid
csusm
mathpics
starcraftvods
TheStrokes
Zombie
scstreams
tagmonkey
WorldofTanks
Homefront
SecretNaughtySC
warpprism
redditforum
rcr
ns2
willertown
ableton
westchesterpa
papercraft
DotA2
BigShot
ebikes
fastfood
eightjay
windsurfing
day9
AtlantaStarcraft
qed
Erotica
descent
AskAmericanReddit
webgl
nethack
Hearthlings
Clarinet
ocaml
AtlantaThrashers
AnimatedGIF
MSP430
alienth
esports
broodwar
starcraft2
Diablo
gaming
AskReddit
diablo2
diablo3
pics
wowguilds
wow
politics
reddit.com
WoWGoldMaking
nflblitz
fluxflashor
FTH
csharp
Shropshire
DontFeartheReposts
Fikevo
redditguild
Tera


Blizzard graph May 2012
leagueoflegends
wowraf
wow
tpl
Transmogrification
funny
wowguilds
broodwar
blizzardallstars
diablo2
diablo3
WoWComics
wowtcg
aww
Denver
Testingoutthecssyo
Depressedrage
AllThings
Ferrah
allthingsprotoss
Diablofun
AngelsAndAirwaves
Infinite


Blizzard graph May 2013
Liquicity
HPC4L
alcohotails
WoWNostalgia
Cinemagraphs
PS5
allthingsprotoss
AskReddit
blizzardallstars
claymore
Baruch
hearthstone
diablo2
diablo3
DotHack
wowtcg
etiquette
doctorwho
talesfromtechsuppo

Future work

- GDELT Events dataset - https://www.gdeltproject.org/
- Dataset containing all significant news since 1970
- Find correlation between Reddit numbers and real world news
- Predict something happened, evaluate the significance of the news

Q&A

Any questions?

nikolaaleksic44@gmail.com
ema.p25@gmail.com