Predicting Engagement:
Social Network Analysis and Deep Learning Techniques

Raffi Mesrobian

Introduction

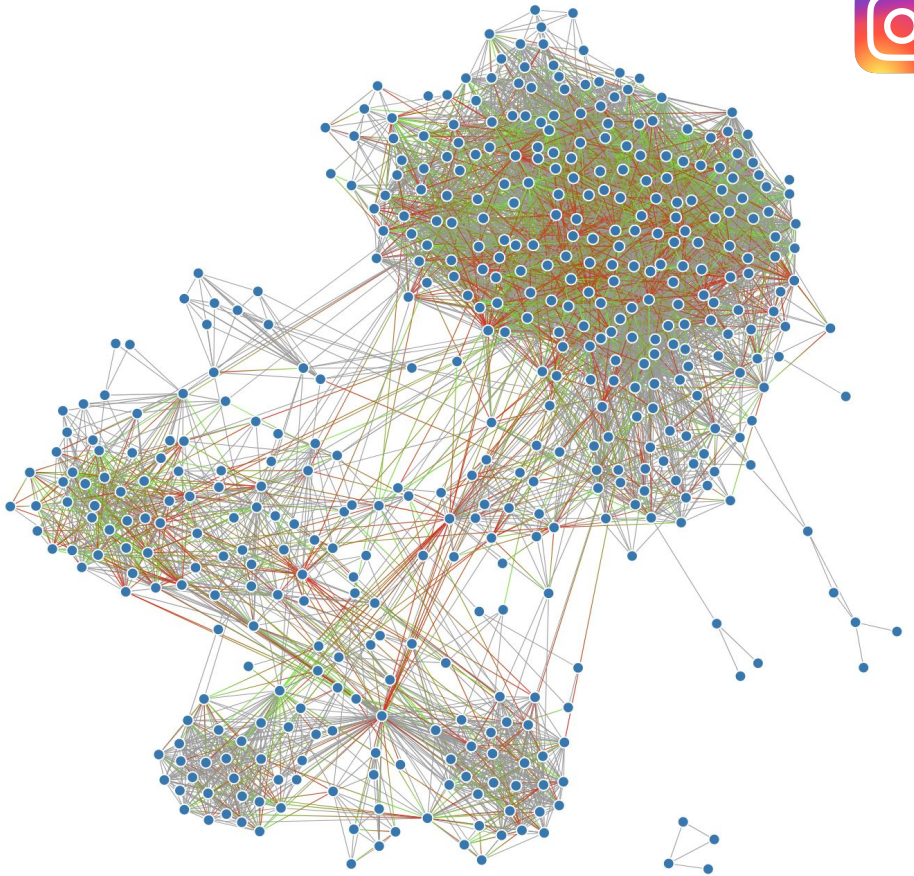
In recent years, social networks have played an increasing role in our daily lives. The social network has become a priority for brand marketing, with Instagram being a platform with one of the highest engagement rates. The data from these networks can be useful for other things, too. This study (1) uses publicly available data from Reddit to examine the effectiveness of different methods that can provide an early detection of major depressive disorders (MDDs).

The objectives of this project are to visualize my social network, applying graph theory to better understand it and to create a model that will predict the engagement level of a post.

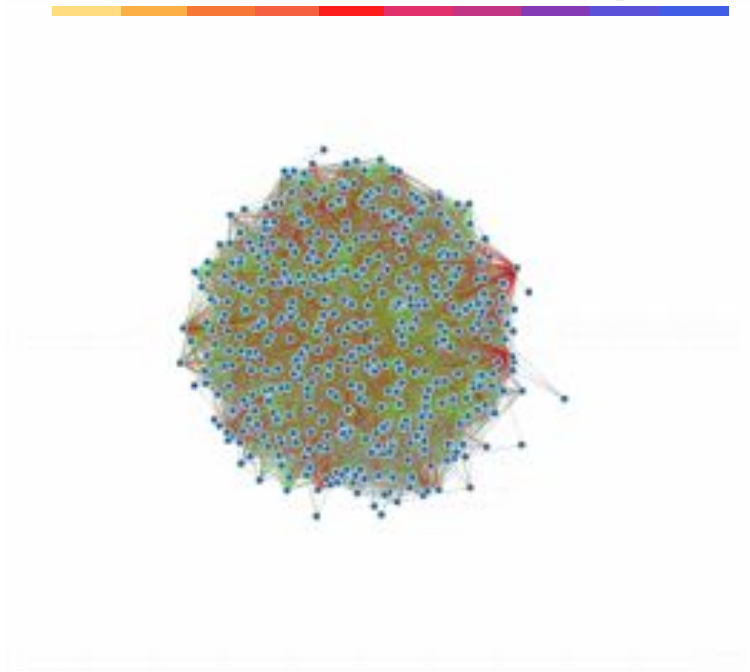
(1) *Cacheda, F, PhD, Fernandez, D, PhD, Novoa, FJ, PhD, Carneira, V, PhD 2019, 'Early Detection of Depression: Social Network Analysis and Random Forest Techniques', JMIR Publications, vol. 21, no 6.*

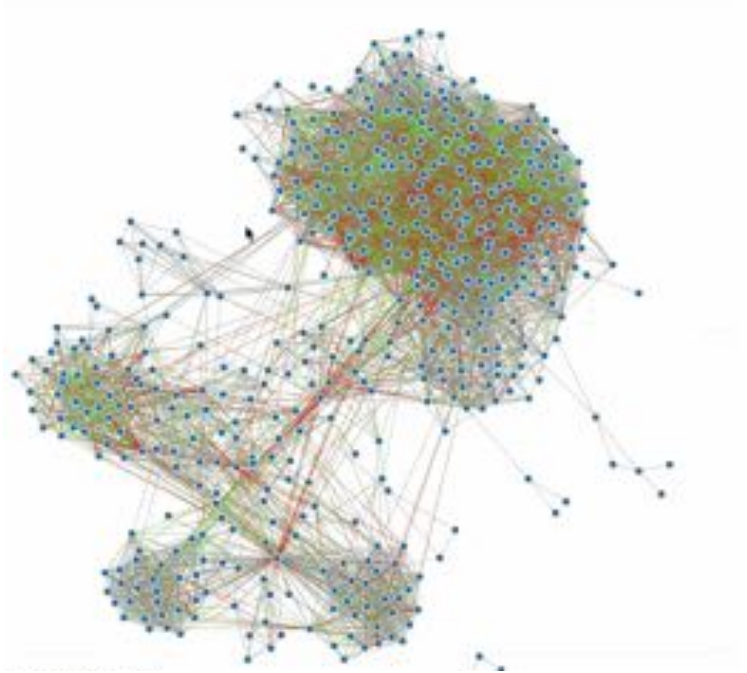
Visualization

- Process:
  - Two web scraping bots
  - txt files to json (d3.js library)
- Network properties:
  - 463 nodes → 11,988 edges
  - Density: 5.6%
  - Avg. Degrees: 25.9
  - Avg. Shortest Path Length: 2.9
- This gist and medium post are great references for those who are interested in visualizing their network and learning about global and local network analysis

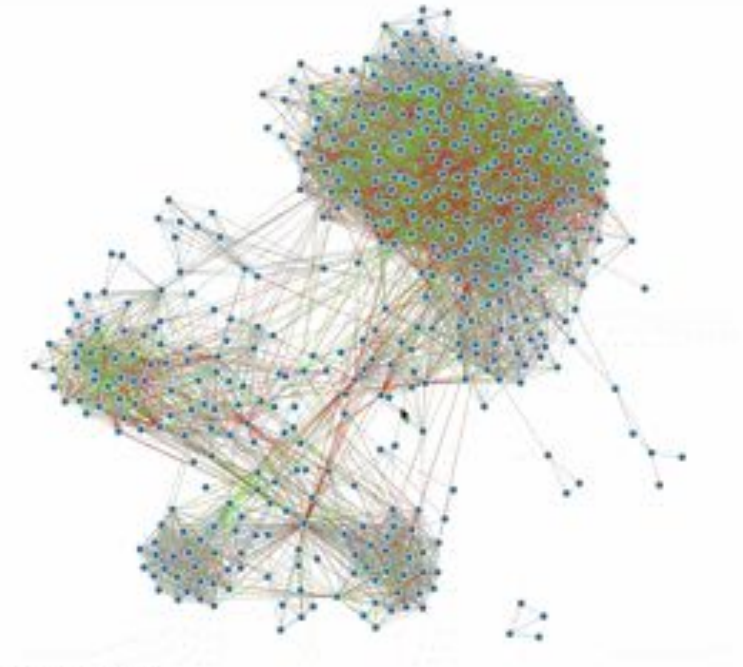


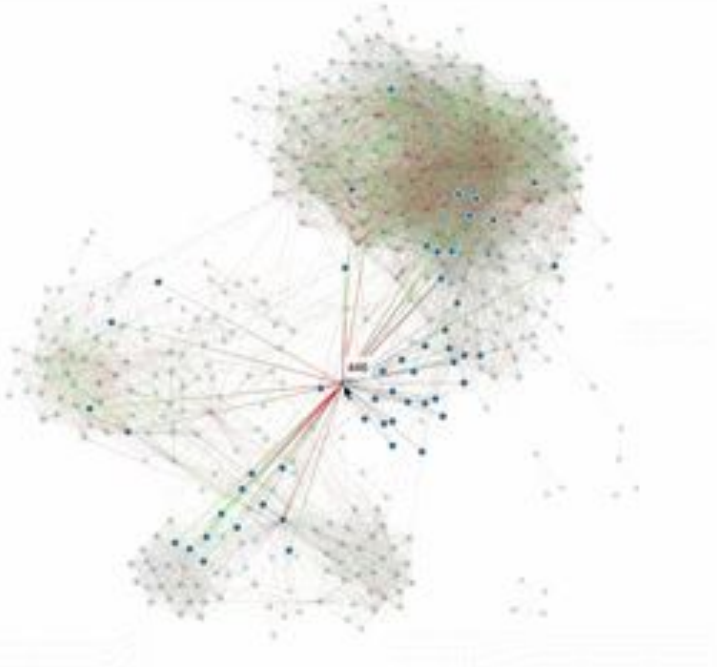
Force Directed Graph



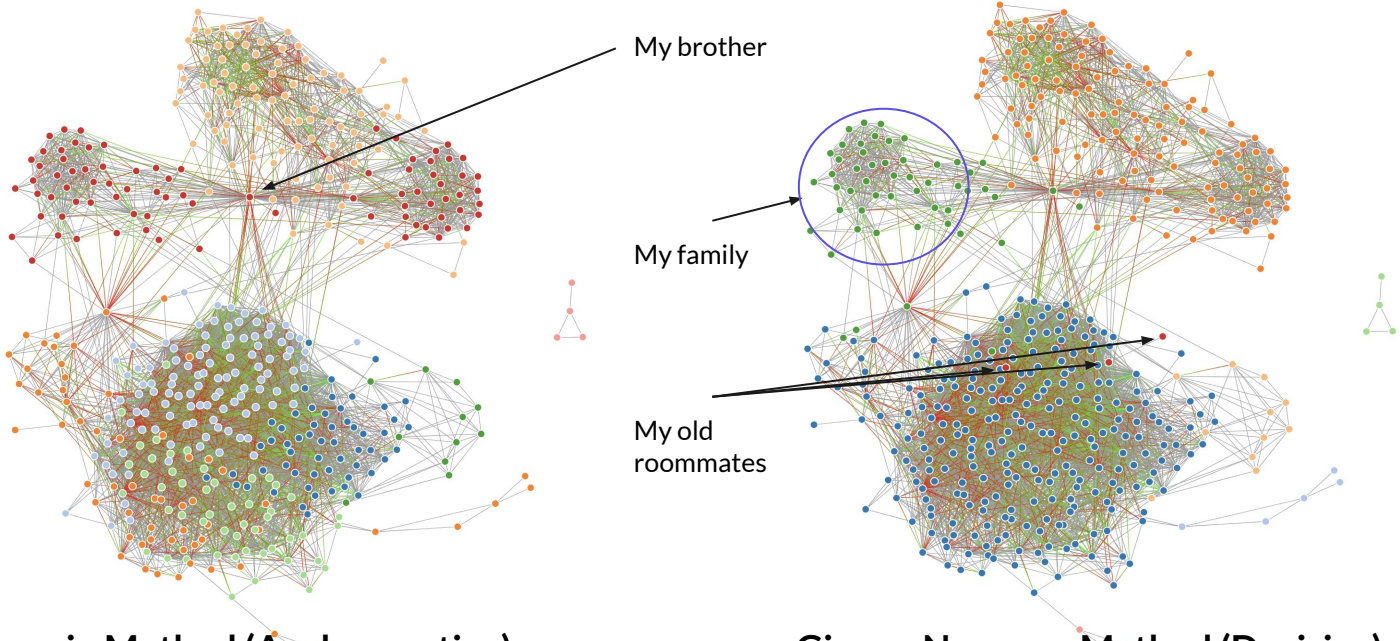


Force Directed Graph (cont.)





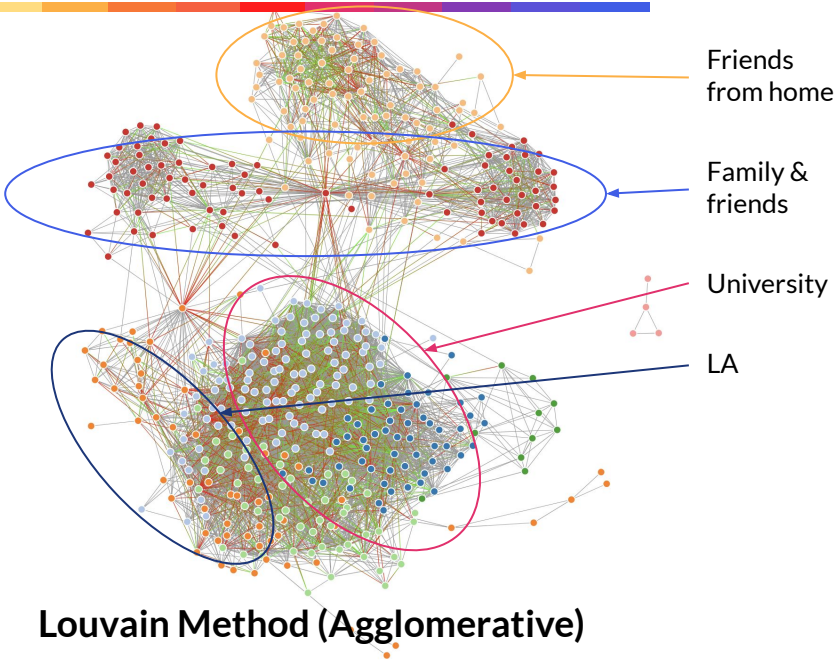
Clustering



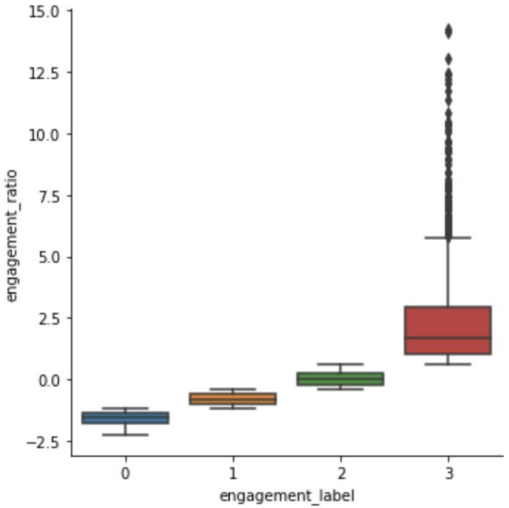
Louvain Method (Agglomerative)

Girvan-Newman Method (Decisive)

Classification



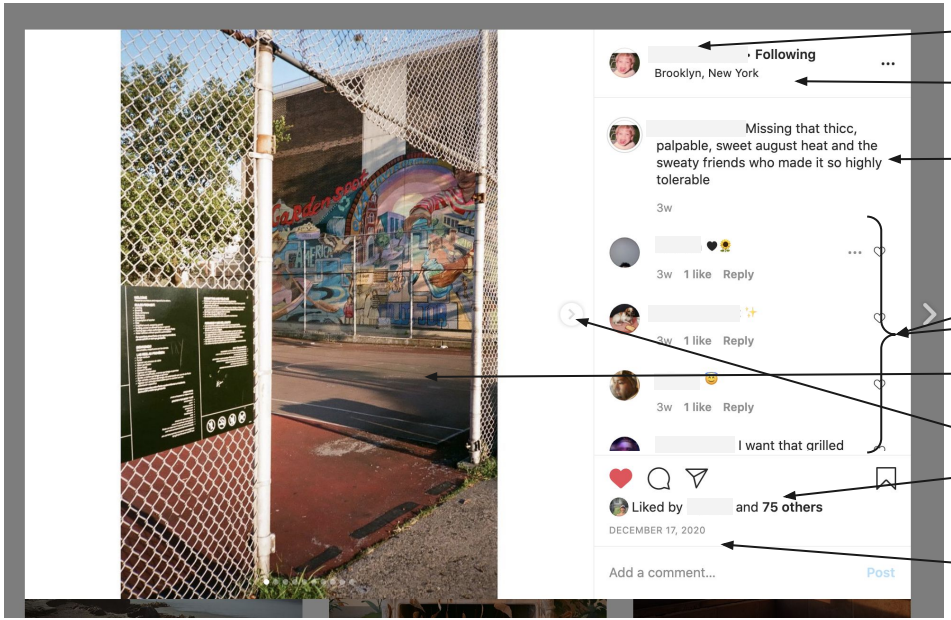
Louvain Method (Agglomerative)



Engagement (0 - 3)

Note: LR = likes / followers; CR = comments / followers; Mean Norm LR (MNL) = (LR - avg. LR) / std. LR; Mean Norm CR (MNC) = (CR - avg. CR) / std. CR; Engagement Ratio = MNL + MNC

Post Data



Username

Location

Caption text

Comment count

Comment text

Image URL

Like count

Type of image

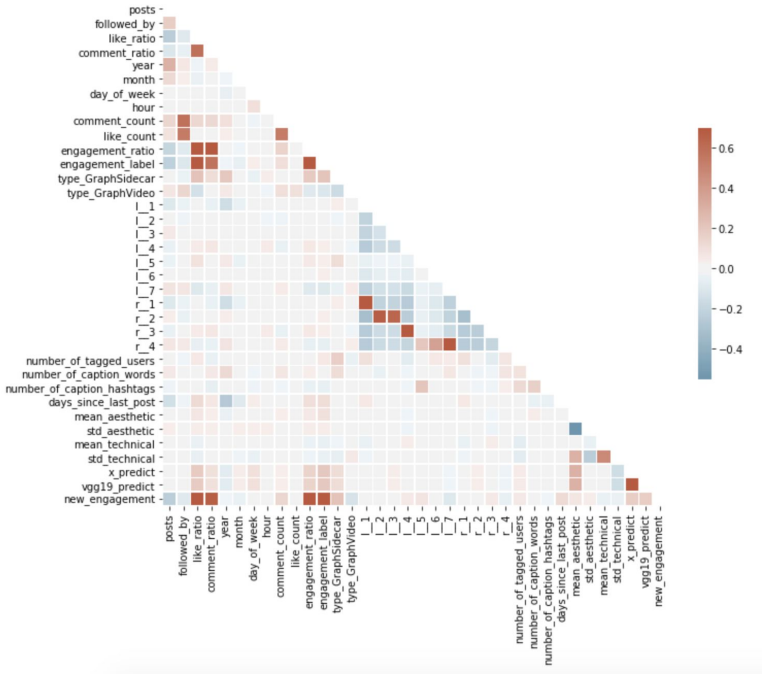
Date and time of post

Classification Challenges

- **Images:**
  - Xception & VGG-19
  - Accuracy: ~33%
  - 😞
- **Natural Language Processing:**
  - Multinomial NB, SGD, Logistic Regression
  - Accuracy: ~45%
  - 😐

Like Ratio:	**0.14**	**0.24**
Comment Ratio:	**0.002**	**0.007**
Caption:	Spent a week in the mountains wondering what month it was. Turns out it's Juntober. Looking forward to Febgust.	**null**
Hashtags:	#undersocialized, #oversaturated	**null**
Community:	**1**	**0**
Engagement Group:	**1**	**2**

Regression

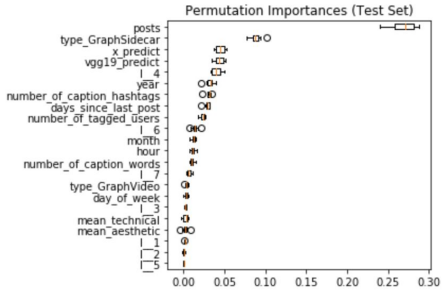


NIMA

Engagement!

Regression (cont.)

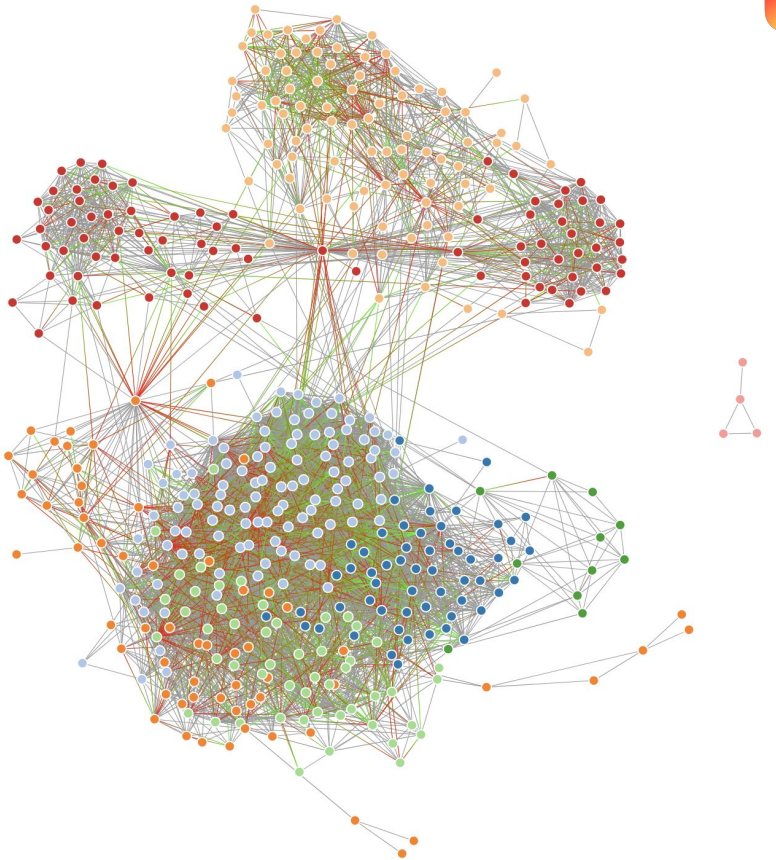
- **Images:**
  - Xception MSE: 3.11
  - VGG-19 MSE: 3.11
- **Tabular Data:**
  - LR MSE: 2.66
  - SVD MSE: 3.20
  - **GB RF MSE: 1.89**



Engagement Ratio:	**-0.775**	**0.484**	**-0.275**
Xception:	-0.009	-0.006	-0.106
VGG-19:	-0.009	-0.006	-0.106
Linear Regression:	-0.086	0.521	1.455
Singular Value Decomposition:	0.012	0.440	1.392
Random Forest:	-0.540	0.640	0.392

Next Steps

- Scrape more data / add more features
- OpenCV + Tesseract → brand logos
- Become an influencer



Thank you!

Xception Image Labels for top 100 most engaging posts:
Siamese_cat, marimba, maypole, Band_Aid, **sunglasses**, Windsor_tie, theater_curtain, stole, candle, Brittany_spaniel, restaurant, Sunglass, suit, paddle, cliff, pajama, cowboy_hat, bikini, coho, studio_couch, ice_lolly, alp, jack-o-lantern, coil, academic_gown, web_site, altar, swing, sunscreen, stage, **hot_pot**, palace, swimming_trunks, seashore, sarong, jean, **burrito**, stingray, ski, **wig**, shower_cap, Pembroke, sombrero, Maltese_dog, breakwater, book_jacket, velvet, **giant_schnauzer**, cannon, seat_belt, rugby_ball, lakeside, valley, bib, pop_bottle, **diaper**, comic_book, cradle, tub, wok, cliff_dwelling, Chesapeake_Bay_retriever, pay-phone, hoopskirt, racer, Australian_terrier, **mobile_home**, slot, croquet_ball, vase, **lab_coat**, maillot, motor_scooter, water_jug, gown, balance_beam, abaya, canoe

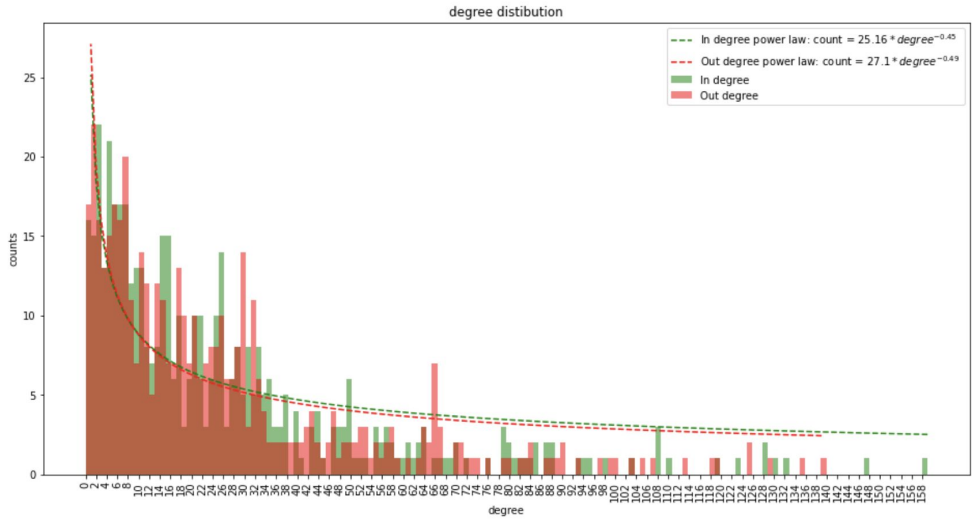
Hashtags in top 100 most engaging posts:
#WhenToomsBecomeOne, **#backstreetboys**, #dnatour2019, #bsbmil, #cincodemacha, #Cindyisababe, #SHS, **#randomactsofkindness,** #motherslove, #motherdaughter, #loveoflife, #loveofmother, #godsmasterpiece, #actsofkindness, #kindness, #kindnessisfree, #kindnessrocks, #mothers, #cousinlove, #sisterslove, #payitforward, **#mariokart**, #mariokart8deluxe, #mariokartpumpkin, #mariopumpkin, #nintendo, #nintendopumpkin, #tanooki, #tanookimario, #sprezzadallas, #tanookipumpkin, #mayorann, #tanookimariopumpkin, #900lumens, #varmit, #racingslicks, #marieloseszinnyc, **#mustardinmyhair**, #hotdogger, #hotdogsog, **#birthdaydog**, #nightmaresquad, #fallofsaigon, #Bozscaggs, #midnighttrain, #secondbank, #fazolis, #alpastor, #1917, #elchapo, #apocalypto, #minne, #gogophs, #bigcat, #ChallengeAccepted, #pgachampionship, #womensupportingwomen, #sacrécoeur#sacrecoeur#frame, #RollTide, #nfl, #oakley, #oakleysunglasses, #theGrandestCanyon, #LTD, #livmichelleconnolly, #baltimoreravens, #ravens, #lamarjackson, #truzz, #bigtruss, #womenempowerment, #acrylicpainting, #acrylic, #acrylicpaint, #football, #portrait, #portraitpainting, #painting, #fantasyfootball, #baltimore, #team43, #worldracingleague, #baltimoremaryland, #contemporaryart, #fineart, #art, #artwork, #latheater, #cantzopthis, #nflmemes, #mvp, #usmc, #wrl, #enduranceracing, #miata, #prettyyoungthing, #wishusluck🙏, #thisisbeyonnnnndme, #felixjohnstone, #munclegiarekracing, #gp4, **#womenempoweringwomen,** #oscarwilde, #devilpup, #teufelswelpe, #hydrate, #gerberMRE, #america, #glendale, #brandpark, #missprism, #MostFunWins, #MarriedMurpy, #TeamSpecialSauce, #Shamrocks, #ScoochButts, #ThankYouEveryone, #adelinafahey

Raffi Mesrobian
rmesrobi
raffimesrobian
rpmesrobian

Appendix

Network Properties: Global Level Analysis

- Density: 5.6%
- Avg. Degrees: 25.9
- Avg. Shortest Path Length: 2.9



Network Properties: Local Level Analysis

Rank 1 - 10

Betweenness Centrality → The bridges

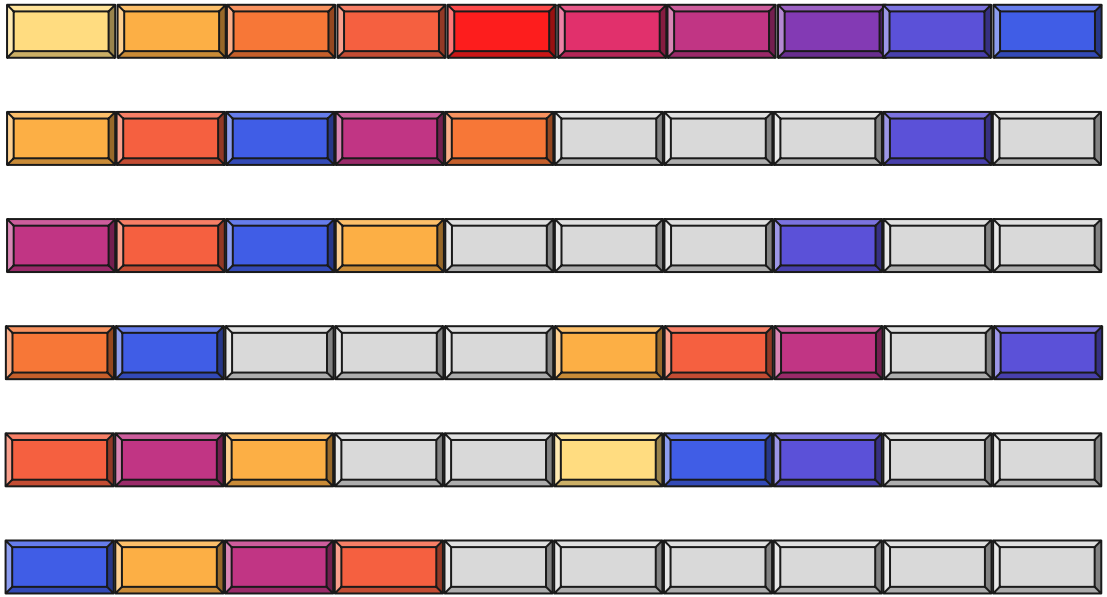
Closeness Centrality → The connectors

In-Degree Centrality → The influencers

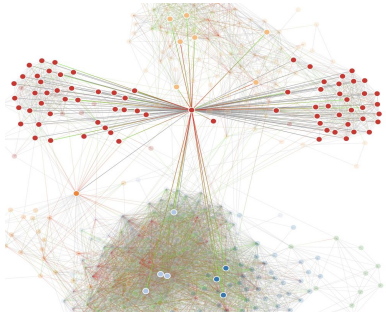
Out-Degree Centrality → The influencers

PageRank → The who knows who

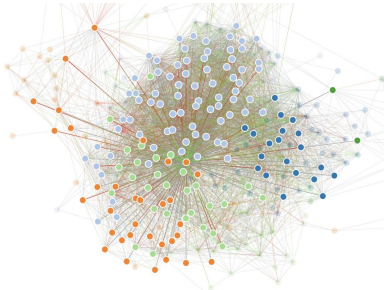
Eigenvector → The who knows who



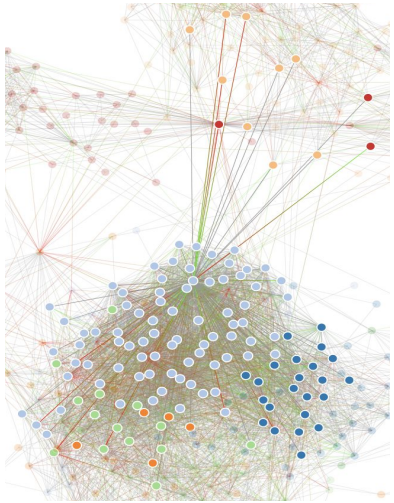
Network Properties: Local Level Analysis (cont.)



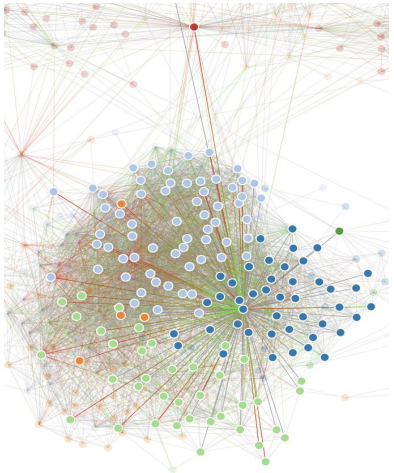
Betweenness Centrality → The bridges



Closeness Centrality → The connectors



In-Degree Centrality → The influencers



PageRank → The who knows who

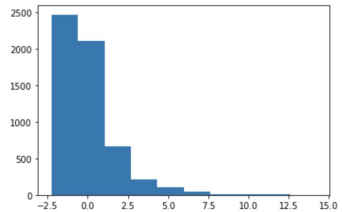
Regression Flow

Engagement Metric:

like_ratio =
likes / follower

mean_norm_likes =
(like_ratio - mean like_ratio) / std like_ratio

engagement_ratio =
mean_norm_likes + mean_norm comments



Feature Engineering & Numerical Data Split:

# of posts
Followed by
Year, month, day, hour
Type of image (OHE)
Community (OHE)
tagged users
of caption words
# of caption hashtags
Days since last post

NIMA Neural Image Assessment:

Mean_aesthetic: 5.846256
Mean_technical: 4.911221

Methods:

Xception,
VGG-19,
Linear Regression,
Random Forest,
Singular Value Decomposition

129,651 data points

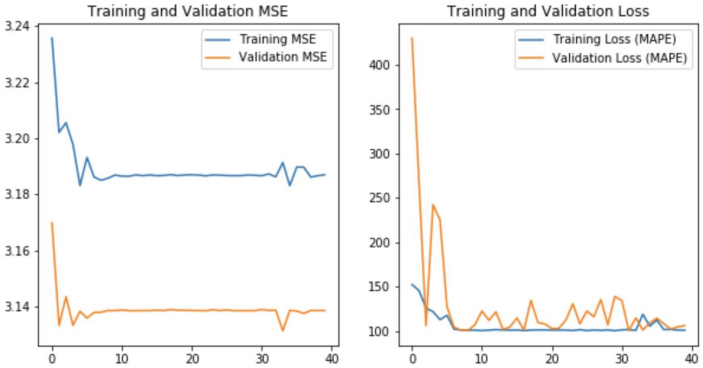
22,548 data points

Train / Test
163,199 points

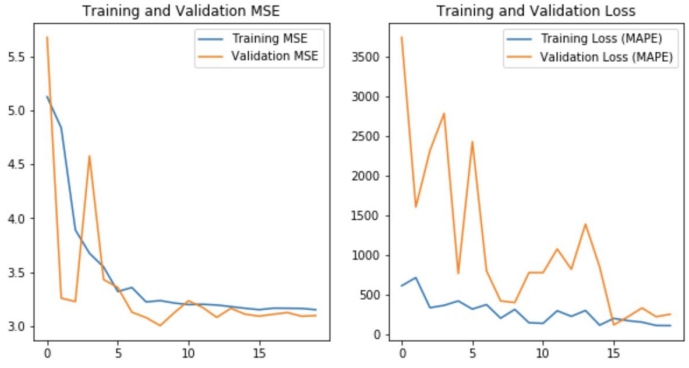
Regression: Image Train / Validate

- 5637 images (4509 train / 1128 validate)
- Image augmentation to help with overfitting

Transfer Learning: VGG-19



Transfer Learning: Xception



Regression Results

Actual ER: -0.275

Features	Method	R2	MAE	MSE	ER
Metadata + Feature Engineering (23 Features)	LR	0.09	1.22	3.16	1.192
	SVR	-0.32	1.57	4.70	2.103
	RF	0.41	1.01	2.21	0.026
Image	**Xception**	**0.02**	**1.22**	**3.11**	**-0.106**
	VGG-19	0.02	1.22	3.11	-0.106
All (27 Features)	LR	0.11	1.17	2.66	1.455
	SVR	0.10	1.21	3.20	1.392
	RF	**0.32**	**0.98**	**1.89**	**0.392**

Logos


aws


BeautifulSoup


K
Keras


matplotlib


NetworkX
Network Analysis in Python


NumPy


pandas


scikit
learn


SciPy


Se
Selenium


seaborn


TensorFlow