Analyzing Influence Metrics in Twitter

Alexandros Dimitrios Keros

February 25, 2016

Introduction

- **Today:**Social Networks redefine the way we communicate
  - Follow, Retweet, Like, Comment, poke, #Hashtag
- **Aim:** Grasp this new sociocultural phenomenon
- **Here:**
  1. Represented Twitter network via graph
  2. Dived into local hashtag-based networks
  3. Attempted to capture percolation
  4. Applied well know centrality measures

Table of Contents

Theoretical Background

1. Graphs
2. Centrality measures
3. Kendall Tau correlation

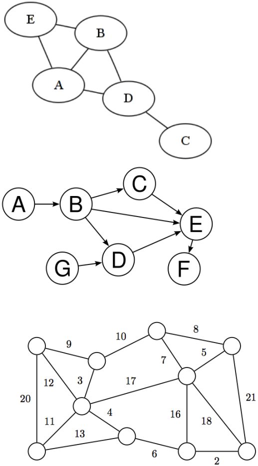
Graphs

$G = (V, E)$

$|V|$ vertices

$|E|$ edges

- Undirected Graphs
- Directed Graphs
- Weighted Graphs



Centrality Measures

1. Degree Centrality
2. Closeness Centrality
3. Betweenness Centrality
4. Eigenvector Centrality
5. PageRank

Degree Centrality

$$C_D(v) = deg(v)$$

- number of nodes that can reach this node *directly*
- focus on number of relations
- reveal local popularity
- *Also*:in-degree centrality, out-degree centrality

Closeness Centrality

$$C_C(v) = \frac{1}{\sum_y d(v, y)}$$

- how fast can a node reach *everyone* in the network
- focus on actor proximity
- reveal communication capacity
- *Also*:weighted closeness centrality

Betweenness Centrality

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

- likelihood of a node being the *most direct route* between other nodes
- focus on intermediary actors
- reveal brokers and privileged actors in information flow
- *Also*:weighted betweenness centrality

Eigenvector Centrality

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{\lambda}\boldsymbol{x}$$

- connection to *other well connected* nodes
- focus on connections of neighbors
- reveal well connected actors
- *Also*:weighted eigenvector centrality

$$PR(v) = \frac{1-d}{N} + d * \sum_{u \in B_v} \frac{PR(u)}{L(u)}$$

d:damping factor
N:number of nodes
L(u):outbound links of node u
B(u):set of pages linking to u

- *eigenvector centrality variant*
- count number and quality of links to a page to determine rough importance

$$\tau = \frac{C - D}{C + D} = \frac{C - D}{\frac{n*(n-1)}{2}}$$

C:concordant pairs
D:discordant pairs
n:sample size

- non-parametric measure of correlation between ranked variables
- probability of difference of the concordant pairs and the discordant pairs
- *p-value*: probability of receiving observed results when *Null-Hypothesis* is true
- *Also*:tau-b, tau-c handle ties

Implementation

- Tools
- Graphs and Centrality measures

Tools

- NetworkX
  Implementations of centrality measures
- scipy.stats
  Kendall Tau Beta implementation

Additional Centrality Measures

- Followers Centrality
- Centralities Euclidean Norm Centrality

$$C_F(v) = followers(v)$$

followers(v):followers of user v

- Reveal popular actors

Centralities Euclidean Norm Centrality

$$C_{EN} = \sqrt{\overline{D_i}^2 + \overline{C_i}^2 + \overline{B_i}^2} \tag{1}$$

$$\overline{D_i} = \frac{D_i - min(\{D_1 ... D_n\})}{max(\{D_1 ... D_n\}) - min(\{D_1 ... D_n\})} \tag{2}$$

$$\overline{C_i} = \frac{C_i - min(\{C_1 ... C_n\})}{max(\{C_1 ... C_n\}) - min(\{C_1 ... C_n\})} \tag{3}$$

$$\overline{B_i} = \frac{B_i - min(\{B_1 ... B_n\})}{max(\{B_1 ... B_n\}) - min(\{B_1 ... B_n\})} \tag{4}$$

D_i:degree centrality score of node i

C_i:closeness centrality score of node i

B_i:betweenness centrality score of node i

Centralities Euclidean Norm Centrality(Cont.)

- Bridge gaps between centrality measures
- Capture both node position and local popularity
- Na Li; Gillet, D., "Identifying influential scholars in academic social media platforms," in Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on , vol., no., pp.608-614, 25-28 Aug.2013

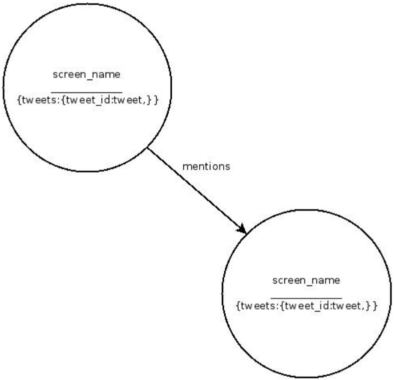
Simple Mentions Graph

$G = (V, E)$

$v \in V$ represents users

$e \in E$ represents mentions

- unweighted directed graph
- Centrality measures applied:
  - in-degree
  - betweenness
  - pagerank
  - eigenvector
  - followers
  - centralities euclidean norm
- **Aim**:Basic network representation



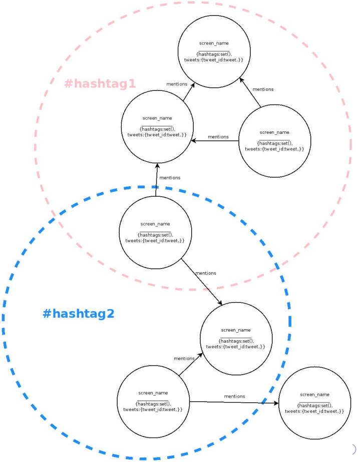
Local Networks Graph

$G = (V, E)$

$v \in V$ represents users

$e \in E$ represents mentions

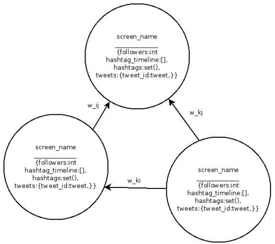
- unweighted directed graph
- Centrality measures applied **both globally and in subgraphs**:
  - in-degree
  - betweenness
  - pagerank
  - eigenvector
  - followers
  - centralities euclidean norm
- **Aim**:Capture local, topic-specific



Weighted Graph

$$G = (V, E)$$
$v \in V$ represents users
$$w_{ij} = wf_j + wh_{ij} + \epsilon$$
$$wf_(j) = \frac{fl(j) - min(\{fl(n_1)...fl(n_n)\})}{max(\{fl(n_1)...fl(n_n)\}) - min(\{fl(n_1)...fl(n_n)\})}$$
$$wh_ij = \begin{cases} \frac{|h_k|}{|hashtags|}, & \text{if j used } h_k \text{ before i} \\ 0, & \text{otherwise} \end{cases}$$
$fl(v)$:followers of user v

- weighted directed graph
- Centrality measures applied:
  - weighted betweenness
  - weighted eigenvector
  - followers
- **Aim**: Capture twitter specific popularity (followers,#hastags)



Results

- Experimental setting and Evaluation methods
- Results

Experimental setting and Evaluation methods

- Dataset: Twitter feed
  - 23956 users
  - 26302 mentions
  - 26302 tweets
- Evaluation methods:
  - Presenting top 5 scorers for each method
  - Kendall tau correlation
  - Computing overlap among 10

Simple Mentions Graph Results - Top 5

In-degree	Betweenness	Pagerank
(civicua, 0.049885)	(guardian, 3380.)	(civicua, 0.014)
(RitaFerrer, 0.027552)	(FGMsilentscream, 1203.)	(Madonna, 0.008)
(Madonna, 0.018076)	(Slate, 585.)	(eonline, 0.007)
(guardian, 0.014527)	(pitchforkmedia, 499.)	(KyivPost, 0.006)
(cnnbrk, 0.014110) beamer@framepauses	(daraobriain, 484.)	(RitaFerrer, 0.006)
Eigenvector	**Followers**	**Cent.Euc.Norm**
(BendyGirl, 0.699456)	(CNN, 11913629.)	(guardian, 1.078)
(sarasiobhan, 0.497058)	(jimmyfallon, 11703692.)	(civicua, 1.)
(Finias, 0.390935)	(UberSoc, 11282590.)	(spaghetti_soup, 1.)
(Jules_Clarke, 0.228532)	(nytimes, 11067872.)	(SociallySavv, 0.842)
(SJaneBernal, 0.162403)	(iamdiddy, 9532092.)	(Derek_Florey, 0.773)

Simple Mentions Graph Results - Kendall Tau

	Betw	PR	Eig	Fol	CEN
InDeg	0.688986	0.978091	0.720850	-0.160175	0.013396
Betw		0.674877	0.919920	0.057288	0.171943
PR			0.712711	-0.167771	-0.000238(p-v:0.95)
Eig				0.019284	0.150419
Fol					0.206929

beamer@framepauses

Simple Mentions Graph Results - Overlaps

	Betw	PR	Eig	Fol	CEN
InDeg	30.03%	84.43%	70.46%	13.32%	0.79%
Betw		20.12%	60.87%	2.10%	1.56%
PR			25.45%	1.45%	0.71%
Eig				0.54%	10.45%
Fol					15.46%

beamer@framepauses

Local Network Graph Results - Top for 3 most frequent hashtags

	In-degree	**Betweenness**
Ukraine	$(\mathrm{RT}_c om, 0.042381)$	(Steiner1776,4.)
Venezuela	(SIGUEMEPRIMERO,0.051969)	(1000riot,0.0)
euromaidan	(SIGUEMEPRIMERO,0.051969)	(1000riot,0.0)
	Eigenvector	**Followers**
Ukraine	(guidestone33,0.809016)	$(\mathrm{RT}_c om, 608697.)$
Venezuela	(hernandezihf,1.)	$(\mathrm{Zapata}_z os, 466705.)$
euromaidan	(hernandezihf,1.)	$(\mathrm{Zapata}_z os, 466705.)$

beamer@framepauses

Local Network Graph Results - Kendall Tau (extreme values)

- Ukraine:
  - max: in degree - pagerank 0.999284
  - min: betweenness - follower 0.038063
- Venezuela:
  - max: in degree - pagerank 0.999938
  - min: eigenvector - follower 0.038063
- euromaidan:
  - max: in degree - pagerank 0.999936
  - min: eigenvector - follower 0.049391

Weighted Graph Results - Top 5

w.Betweenness	w.Eigenvector	Followers
(guardian ,3380.)	(BendyGirl, 0.699456)	(CNN, 11913629.)
(FGMsilentscream ,1203.) beamerCframepauses	(sarasiobhan, 0.497058)	(jimmyfallon, 11703692
(Slate ,585.)	(Finias, 0.390935)	(UberSoc, 11282590.
(pitchforkmedia ,499.)	(Jules$_{C}$*larke*, 0.228532)	(nytimes, 11067872.)
(daraobriain ,484.)	(SJaneBernal, 0.162403)	(iamdiddy, 9532092.)

Weighted Graph Results - Kendall Tau & Overlaps

beamer@framepauses

	Eig	**Fol**
Betw	0.919920,85.87%	0.057288,1.10%
Eig		0.019284,0.79%

Conclusions

- Different centrality methods reveal a different aspect of the network
- Network representation is important

Future Work

- Topic extraction (Latent Dirichlet Allocation e.t.c)
- Percolation simulation and centrality (i.e. take network evolution into account)
- More sophisticated network representation (Hypergraphs)
- Influence sources are often subjective