Duanbing Chena, Linyuan Lüb, Ming-Sheng Shanga, Yi-Cheng Zhanga, Tao Zhoua,

IDENTIFYING INFLUENTIAL NODES IN COMPLEX NETWORKS (2011)

Presented by Carles Sans Fuentes

Linköpings universitet

Social Network Analysis

THE FIRST RELLEVANT STUDY ABOUT DEGREES OF SEPARATION ON SOCIETY DATES FROM 1960

The Small World experiment (1960)

SET UP

Objective

Average path length for social networks of people in the US

Instructions

Send to "randomly" selected individuals (from Nebraska or Kansas) information packets:
- study's purpose
- target contact person
- Rules

Main Rule:
If(knew person == TRUE){
forward the letter directly
} Else {
think of a friend or who was more likely to know the target
}

RESULTS

Reaching the destination

- 78% didn't reach destination (232 of the 296 letters)
  - ✓ large % rejection of on participation
- **22%** (64 of the letters) eventually **did reach the target contact**

CONCLUSIONS

Average path length

6 degrees of separation

Hub people

50% letters from **Kansas** had the **same two people as last contact**

The average degree of separation between people was determined to be 6

Sources:
(1) Milgram, Stanley (May 1967). "The Small World Problem". Psychology Today. Ziff-Davis Publishing Co.
(2) Blass, T. (2004). The Man Who Shocked the World: The Life and Legacy of Stanley Milgram. ISBN 0-7382-0399-8

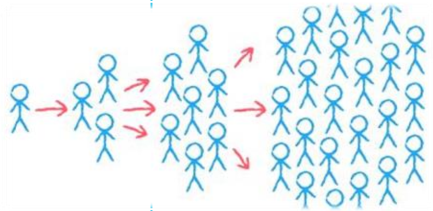
DURING THE LAST DECADES THE DEGREES OF SEPARATION AMONG INTERNET USERS HAS SHRUNKED BETWEEN 3 AND 5

Main facts about social network connections through the last decade

2005 | 2010 | 2015 | 2020

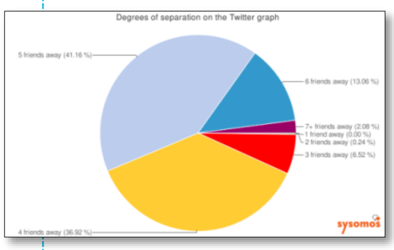
Microsoft

6,6 steps was the average path length among **Messenger** users in **Microsoft** **(2008)** (1)



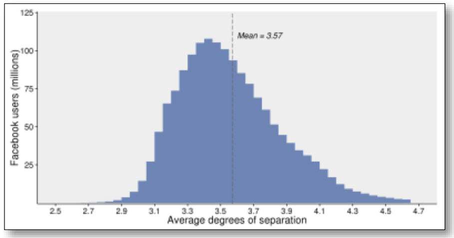
twitter

4,67 steps are the most common friendship distance in **Twitter** **(2010)** (1)



facebook

3,57 steps are the average on how a person is connected to every other person in **Facebook** **(2016)** (2)



Sources:
(1) https://sysomos.com/inside-twitter/twitter-friendship-data/ ;
http://barnraisersllc.com/2012/04/studies-social-media-6-degrees-of-separatio/
(2) https://research.fb.com/three-and-a-half-degrees-of-separation/

UNDERSTANDING SOCIAL NETWORKS IS CRUCIAL TO EVALUATE HUBBS AND THE POWER OF ITS NODES

Top 10 most influential Man

10 @KevinHart4real
9 @BillGates
8 @jimmyfallon
7 @cristiano
6 @realdonaldtrump
5 @zaynmalik
4 @kanyewest
3 @jtimberlake
2 @BarackObama
1 @justinbieber

Top 10 most influential Women

10 @Adele
9 @britneyspears
8 @TheEllenShow
7 @HarukaN_JKT48
6 @ddlovato
5 @Oprah
4 @shakira
3 @selenagomez
2 @katyperry
1 @JLo

By Brandwatch 2016 (1)

Top 10 influential channels

Ranking	Influencer	Likes	Comments	Engagement
1	kimkardashian	443,413,766	5,613,461	449,027,227
2	kyliejenner	383,667,470	33,235,937	416,903,407
3	mileycyrus	379,865,123	5,594,090	385,459,213
4	arianagrande	349,941,016	5,353,575	355,294,591
5	justinbieber	302,037,886	8,909,224	310,947,110
6	nickiminaj	301,564,363	4,707,143	306,271,506
7	khloekardashian	286,570,015	3,478,218	290,048,233
8	neymarjr	266,899,059	19,902,550	286,801,609
9	taylorswift	239,644,922	3,364,992	243,009,914
10	selenagomez	230,039,262	3,435,666	233,474,928

By Instagram 2015(2)

Top 12 influential channels by Subscribers

RANK	SB SCORE		USER	SUBSCRIBERS	VIDEO VIEWS
1	13786949	--	Music	97,068,174	--
2	12893138	--	Gaming	77,716,341	--
3	11859803	--	Sports	75,546,769	--
4	46	A+	PewDiePie	54,914,225	15,203,162,864
5	236069	B-	YouTube Movies	47,121,447	1,006,181
6	13492087	--	News	33,812,289	--
7	899	A	HolaSoyGerman.	31,641,863	3,081,272,681
8	11786124	--	Popular on YouTube	29,665,991	--
9	42	A+	JustinBieberVEVO	28,825,980	15,060,800,995
10	16602	B	YouTube Spotlight	25,511,130	1,100,925,024
11	157	A	RihannaVEVO	24,623,628	11,183,861,175
12	113	A	elrubiusOMG	24,215,352	5,307,366,565

By SocialBlade 2016(3)

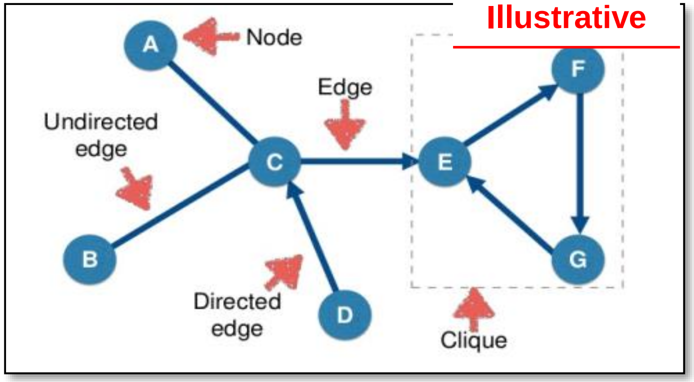
Sources:
(1) https://www.brandwatch.com/blog/react-the-most-influential-men-and-women-on-twitter/
(2):http://www.edigitalagency.com.au/instagram/top-most-popular-instagram-influencers-video-creators/
(3) https://socialblade.com/youtube/top/100/mostsubscribed

A NETWORK GRAPH IS A COLLECTION OF ENTITIES, EACH CALLED A VERTEX OR NODE

Main characteristics of a network/graph

		Definition	Examples
Network (graph)	**Vertex/ Node**	> Collection of entities, each called a vertex or node	> Vertices are mathematicians, edges represent coauthorship relationships > Vertices are Facebook users, edges represent Facebook friendships
	Edges/ Links	> A list of pairs of vertices that are neighbors, representing edges or links	
Type of edges	**Directed**	> Edges have a direction associated with them	
	Undirected	> Edges does not have a direction associated Bidirecttional	
	Weighted	> Edges have a weight associated with them	
	Unweighted	> Edges does not have a weight associated Bidirecttional	

Illustrative



Networks can represent any binary relationship over individuals

Sources:
(1) http://www.cis.upenn.edu/~mkearns/NetworkedLifeOnline/ → what is a network ?
(2) https://www.slideshare.net/bodacea/network-analysis-lecture

INFLUENTIAL NODES ARE AFFECTED BY MECHANISMS SUCH AS CASCADING, SPREADING AND SYNCHRONIZING

Cascading

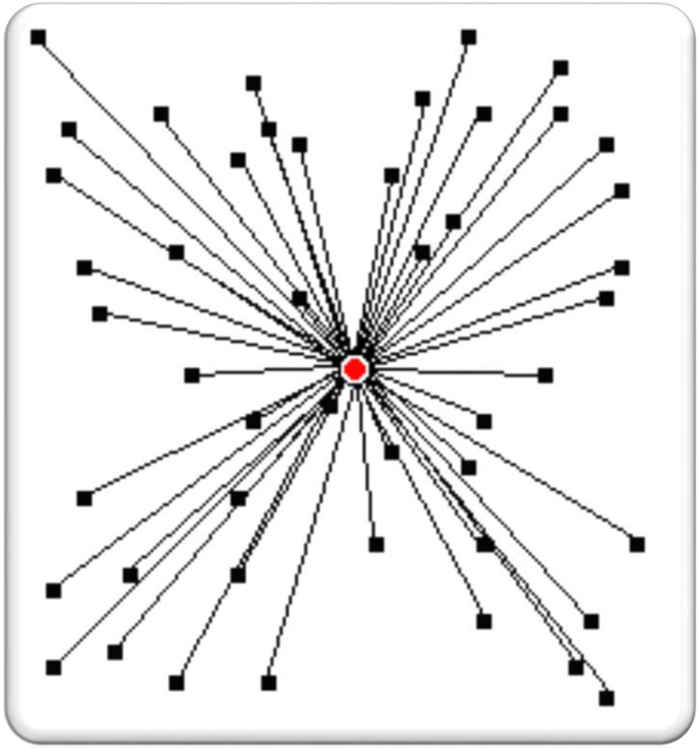
When a person observes the actions of others and then engages in the same acts[1]

Spreading

How information is transmitted to new nodes from an initial node[2]

Synchronizing

Nodes that are able to transmit and receiving information with regular frequency[3]



Identifying influential nodes is of great importance:
e.g. controlling rumor and disease spreading, and creating new marketing tools

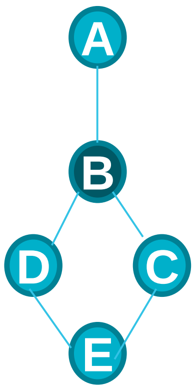
Sources:
(1) https://en.wikipedia.org/wiki/Information_cascade
(2) https://en.wikipedia.org/wiki/Spreading_activation
(3) https://www.techopedia.com/definition/13390/synchronization-dot-net

DEGREE, CLOSENNESS AND BETWEENNES CENTRALITY ARE MEASURES USED TO IDENTIFY INFLUENTIAL CENTRAL NODES

Main centrality measures

	Definition	Example
Average degree of a network[1]	> Average number of steps along the shortest paths for all possible pairs of network nodes $l_G = \frac{1}{n \cdot (n-1)} \cdot \sum_{i \neq j} d(v_i, v_j)$	> $C_c(b)$ = d(*b*, *a*) + *d*(*b*, *c*) + *d*(*b*, *d*) + *d*(*b*, *e*)) / 4) = (1 + 1 + 1 + 2) = 5/4 = 1.2
Degree Centrality	> The number of neighbours a node has (e.g. the number of links it has) O(n^2) $C_D(v) = \deg(v)$	> $C_D(b)$ = A, C, D /A, C,D, E = 3/4 = 0.75
Betweenness Centrality	> Fraction of shortest paths between node pairs that pass through the node of interest. O(n^3) $C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}},$	> $C_b(b) = ((\sigma_{ac}(b) / \sigma_{ac}) + (\sigma_{ad}(b) / \sigma_{ad}) + (\sigma_{ae}(b) / \sigma_{ae}) + (\sigma_{cd}(b) / \sigma_{cd}) + (\sigma_{ce}(b) / \sigma_{ce}) + (\sigma_{de}(b) / \sigma_{de})) / 6$ = ((1 / 1) + (1 / 1) + (2 / 2) + (1 / 2) + 0 + 0) / 6 = 3.5 / 6
Closeness Centrality	> The reciprocal of the sum of geodesic distances to all other nodes of V $C_C(v) = \frac{1}{\sum_{t \in V \setminus v} d_G(v,t)},$ O(n^2*⟨k⟩) ⟨k⟩ is the average degree of the network	> $C_c(b)$ = 1/ ((*L*(*b*, *a*) + *L*(*b*, *c*) + *L*(*b*, *d*) + *L*(*b*, *e*)) / 4) = 4/ (1 + 1 + 1 + 2) = 4/5 = 0.8

Illustrative



Better at **quantifying** the **influence** of nodes, **but higher computational complexity**[2]

It exists a tradeoff between the algorithm of centrality used and the computational complexity

LINKÖPINGS UNIVERSITET

(1) Average degree of a network is not itself a measure of centrality but to understand the article
(2) Centralities based on PageRank or LeaderRank are even more relevant, but more complex
Sources: http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/

THE LOCAL CENTRALITY MEASURE CONSIDERS BOTH THE NEAREST NEIGHTBORS AND THE NEXT NEAREST NEIGHBORS

Main Issue

Finding better algorithm with less complexity

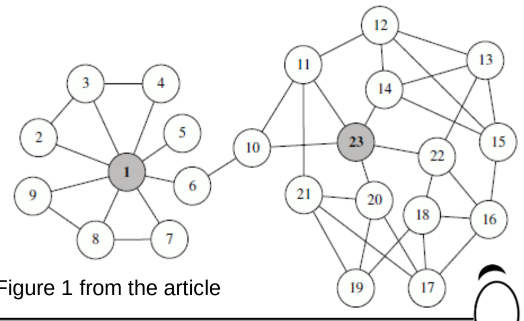


Figure 1 from the article

Degree centrality problem: The network consisted of 23 nodes and 40 edges. Although node 23 has lower degree than node 1, its influence may be even higher

Solution model proposed

Local centrality measure

- It considers both the nearest and the next nearest neighbors
- The local centrality CL(v) of node v is defined as:

$$Q(u) = \sum_{w \in \Gamma(u)} N(w),$$

where Γ (u) is the set of the nearest neighbors of node u and N(w) is the number of the nearest and the next nearest neighbors of node w. Then:

$$C_L(v) = \sum_{u \in \Gamma(v)} Q(u),$$

Example from Figure 1

N(1) = 9
Q(1) = N(2) + N(3) + N(4) + N(5) + N(6) + N(7) + N(8) + N(9) = 67

$C_L(1)$ =
Q(2)+Q(3)+Q(4)+Q(5)+Q(6)+Q(7) + Q(8) +Q(9) = 145

Theoretical advantages

- ✓ Tradeoff between low-relevant degree centrality and other time-consuming measures
- ✓ O(n*⟨k⟩^2) which grows linearly with the size of a sparse network
- ✓ It considers two neighbors level for the importance of spreading

THE SIR MODEL IS AN EPIDEMIC MODEL USED TO EXAMINE THE SPREADING INFLUENCE OF TOPRANKED NODES

	#1 Susceptible	#2 Infection	#3 Recovery
	S(t)	I(t)	R(t)
Typical model information	Nº of individuals not yet infected but susceptible to get it • Beta: % susceptible-infected contact results in a new infection	Nº of individuals who have been infected capable of spreading it • Gamma: % infected recovers & moves into the resistant phase	Nº of individuals who were infected and then removed from the disease (due to immunization or death) .
Main formulas	$\frac{dS}{dt} = -\beta\frac{SI}{N}$	$\frac{dI}{dt} = \beta\frac{SI}{N} - \gamma I$	$\frac{dR}{dt} = \gamma I$
General	$N = S + I + R$		
Assumptions of the model	At each step, for each infected node, one randomly selected susceptible neighbor gets infected with probability Beta = 1/# Neighbours	Infected nodes recover with probability 1/⟨k⟩ at each step, with ⟨k⟩ is the average degree of the network Gamma = 1/⟨k⟩	-

- ✓ An infected node will in average contact ⟨k⟩ neighbors before he/she is recovered
  - The process stops when there is no infected node

- ✓ The total number of infected and recovered nodes at time t, denoted by F (t), can be considered as an indicator to evaluate the influence of the initially infected node at time t
  - higher F (T_c) indicates a larger influence

Sources:
(1) http://www.public.asu.edu/~hnesse/classes/sir.html?Alpha=1&Beta=0.2&initialS=1000&initialI=10&initialR=3&iters=10
(2) https://en.wikipedia.org/wiki/Epidemic_model#The_SIR_model
(3) http://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model

THE OBJECTIVE OF THE ARTICLE IS TO EVALUATE THE PERFORMANCE OF THE LOCAL CENTRALITY MEASURE

Objective

- **Evaluate** the **performance** of **our centrality measure compared with** the **other** centrality measures

Data

i. Blogs—the communication relationships between owners of blogs on the MSN (Windows Live) Spaces website

ii. Netscience—the network of co-authorships between scientists who are themselves publishing on the topic of networks.

iii. Router level topology of the Internet, collected by the Rocket fuel Project

iv. Email—the network of e-mail. Interchanges between members of the University Rovira i Virgili (Tarragona)

Model of evaluate

- The SIR model from the previous slide

Order of performance

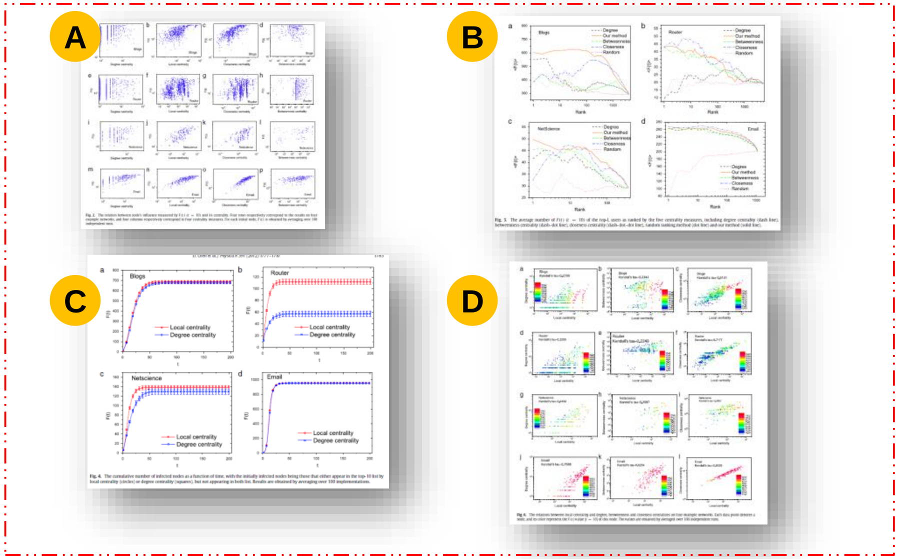
0 Data Features[1]

Network	n	m	$\langle k \rangle$	k_{max}	C	$\langle d \rangle$	r	H
Blogs	3982	6803	3.42	189	0.1409	6.227	−0.1330	4.038
Netscience	379	914	4.82	34	0.3706	6.061	−0.0817	1.663
Router	5022	6258	2.49	106	0.0058	6.393	−0.1384	5.503
Email	1133	5451	9.62	71	0.1101	3.716	0.0782	1.942

1 Methodology

- After n implementations (each node is selected to be the initially infected node once and only once)
- Evaluate the total number of infected and recovered nodes at time t, denoted by F (t) where t = 10 , can be considered as an indicator to evaluate the influence of the initially infected node at time t.

2 Results[2]



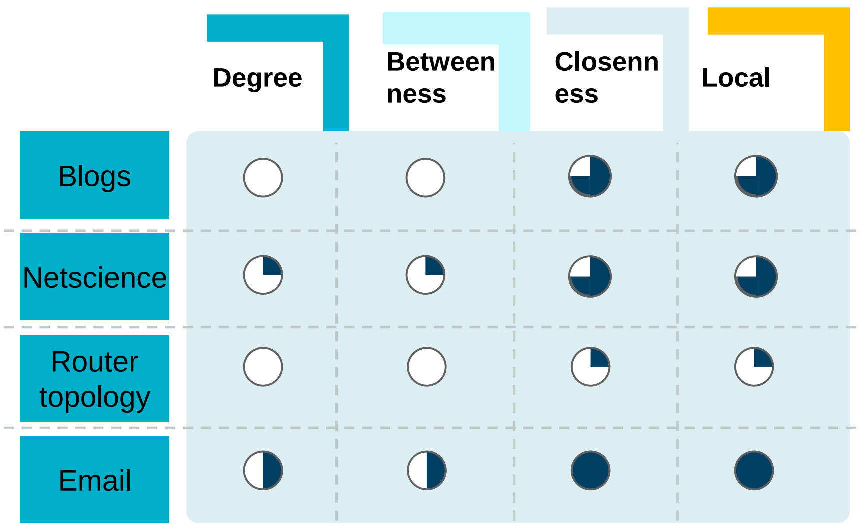
Focus in next slides

(1) n and m are the total numbers of nodes and links, respectively. ⟨k⟩ and kmax denote the average and the maximum degree. ⟨d⟩ is the average shortest distance. C and r are the clustering coefficient and assortative coefficient respectively. H is the degree heterogeneity

(2) Figures from the results and tables with the main characteristics of the networks can be found in big with its description in the annex

THE LOCAL CENTRALITY MEASURE PERFORMS COMPETITIVELY BETTER THAN THE OTHER MEASURES (I)

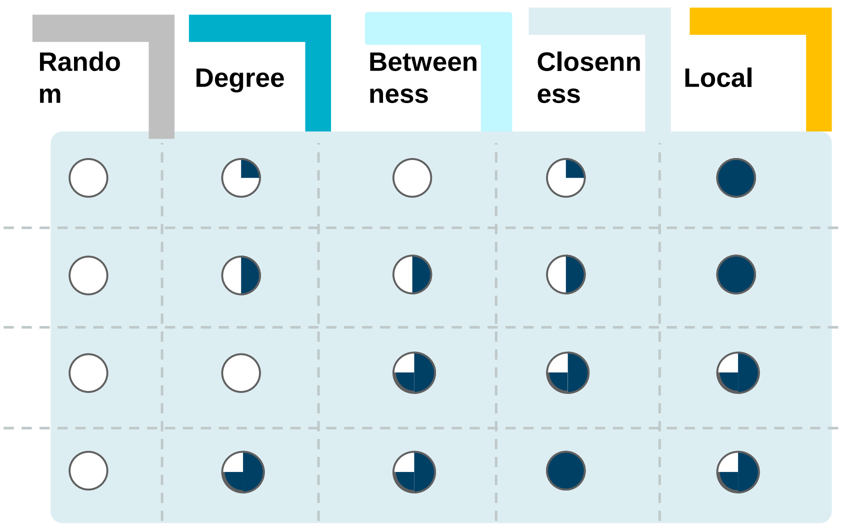
A The relation between node's influence measured by F (t) (t = 10) and its centrality



Correlation between centrality with F (t) : ○ No correlation ● High correlation

✓ Local and closeness centrality measures perform good
X Degree and betweenness centrality perform quite bad

B The average number of F (t) (t = 10) of the top-L users as ranked and its centrality[1]



Performance between Rank with F(t) : ○ Bad ● Excellent

✓ Local centrality measure outperforms
- In Blogs and Netscience local centrality performs best
- In Router and Email closeness centrality performs a little bit better

X Random, Degree and Betweenness centrality perform quite bad

(1) A good centrality measure should be downward sloping: namely the average number of infected nodes by the top-L nodes decreases with the increasing of L.

THE LOCAL CENTRALITY MEASURE PERFORMS COMPETITIVELY BETTER THAN THE OTHER MEASURES (II)

C The cumulative number of infected nodes as a function of time[1]

	Email	Router topology	Netscience	Blogs
Degree				
Local	✓	✓	✓	✓

✓ Local centrality performs better in all cases

D Relations between local centrality and degree, betweenness and closeness[2]

Local	**Degree**	**Betweenness**	**Closenness**
Blogs	$\tau = 0.28$	$\tau = 0.23$	$\tau = 0.61$
Netscience	$\tau = 0.45$	$\tau = 0.26$	$\tau = 0.47$
Router topology	$\tau = 0.23$	$\tau = 0.23$	$\tau = 0.72$
Email	$\tau = 0.76$	$\tau = 0.53$	$\tau = 0.86$

Correlation with Local measure: ○ Bad ● Excellent
Kendals tau used to evaluate concordancy thorugh data

✓ Closeness is strongly positively correlated with local centrality
- In E-mail, all centralities are positively correlated with the local

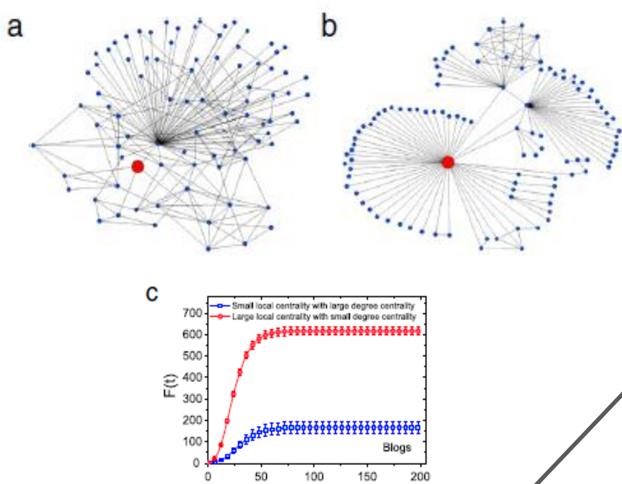
(1) with the initially infected nodes being those that either appear in the top-10 list by local centrality (circles) or degree centrality (squares)
(2) Each data point denotes a node, and its color represent the F (t) value (t = 10) of this node. The values are obtained by averaged over 100 independent runs.

THE LOCAL CENTRALITY PERFORMS BETTER THAN DEGREE CENTRALITY BEING LESS COMPLEX AND MORE EFFECTIVE

Main conclusions of the paper

Local Centrality

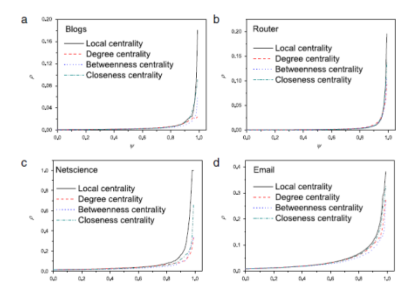
Captures better the F(t) than degree centrality



Performs competitively really good
Better than centrality basic measures

Less complex and More effective

Ranking-based rich-club phenomenon
Good at finding hidden relationships between structure and function of networks



Need of comparison to more complex and efficient measures

Better explanation on:

- 4 networks used
- Ranking-based rich-club phenomenon

Main Conclusions

Criticism

Annex

Relation between node’s influence measured by f (t) and its centrality

Average number of f (t) (t = 10) of the top-l users as ranked and its centrality

The cumulative number of infected nodes as a function of time

Relations between local centrality and degree, betweenness and closeness

Table (iii) top-10 ranked nodes by local centrality

Table (iv) mean value of the top nodes

A ANNEX - RELATION BETWEEN NODE'S INFLUENCE MEASURED BY F (T) (T = 10) AND ITS CENTRALITY

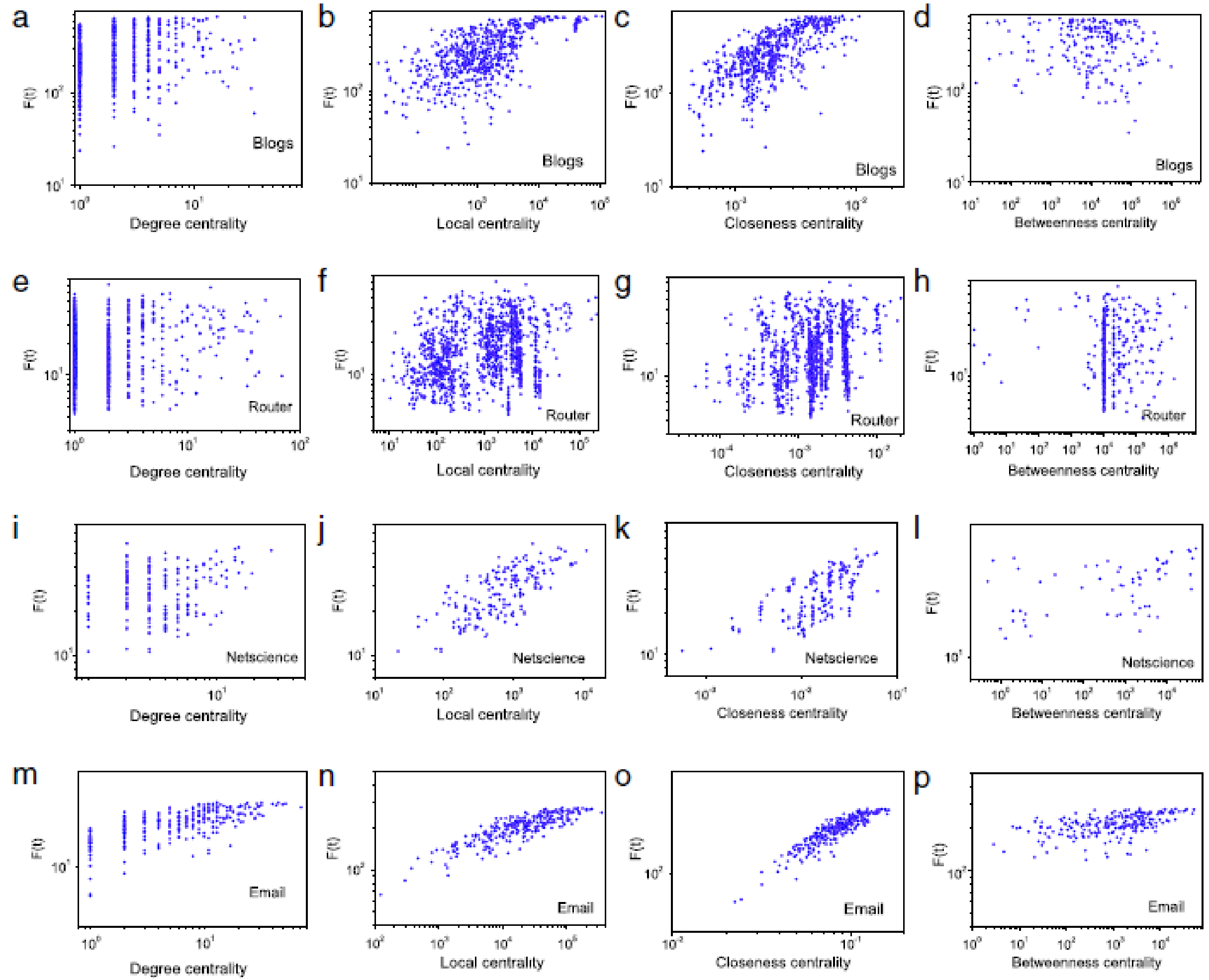


Fig. 2. The relation between node's influence measured by $F(t)$ ($t = 10$) and its centrality. Four rows respectively correspond to the results on four example networks, and four columns respectively correspond to four centrality measures. For each initial node, $F(t)$ is obtained by averaging over 100 independent runs.

Back to results

B ANNEX - AVERAGE NUMBER OF F (t) (t = 10) OF THE TOP-L USERS AS RANKED AND ITS CENTRALITY

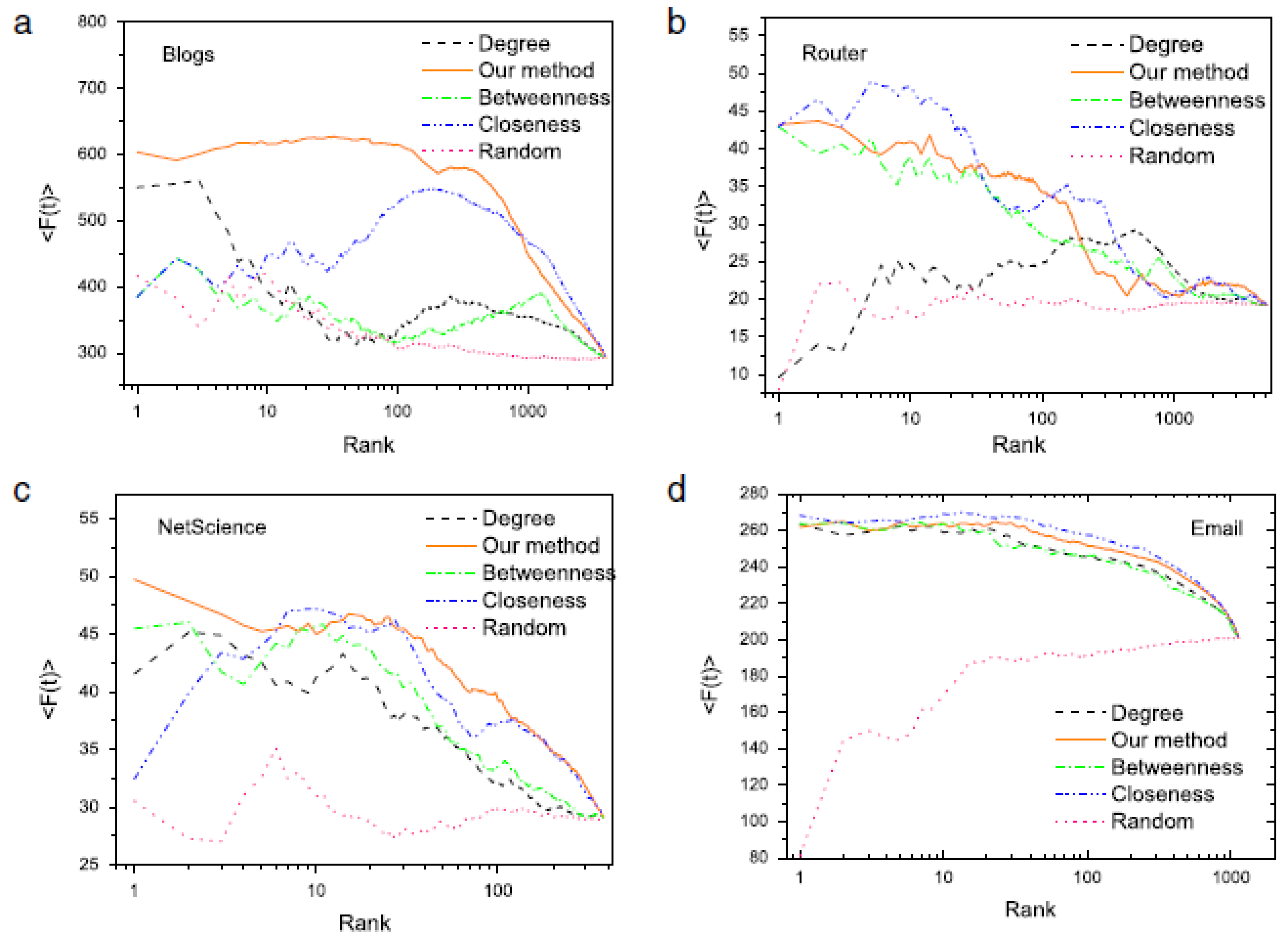


Fig. 3. The average number of $F(t)$ $(t = 10)$ of the top-L users as ranked by the five centrality measures, including degree centrality (dash line), betweenness centrality (dash–dot line), closeness centrality (dash–dot–dot line), random ranking method (dot line) and our method (solid line).

Back to results

C ANNEX - THE CUMULATIVE NUMBER OF INFECTED NODES AS A FUNCTION OF TIME

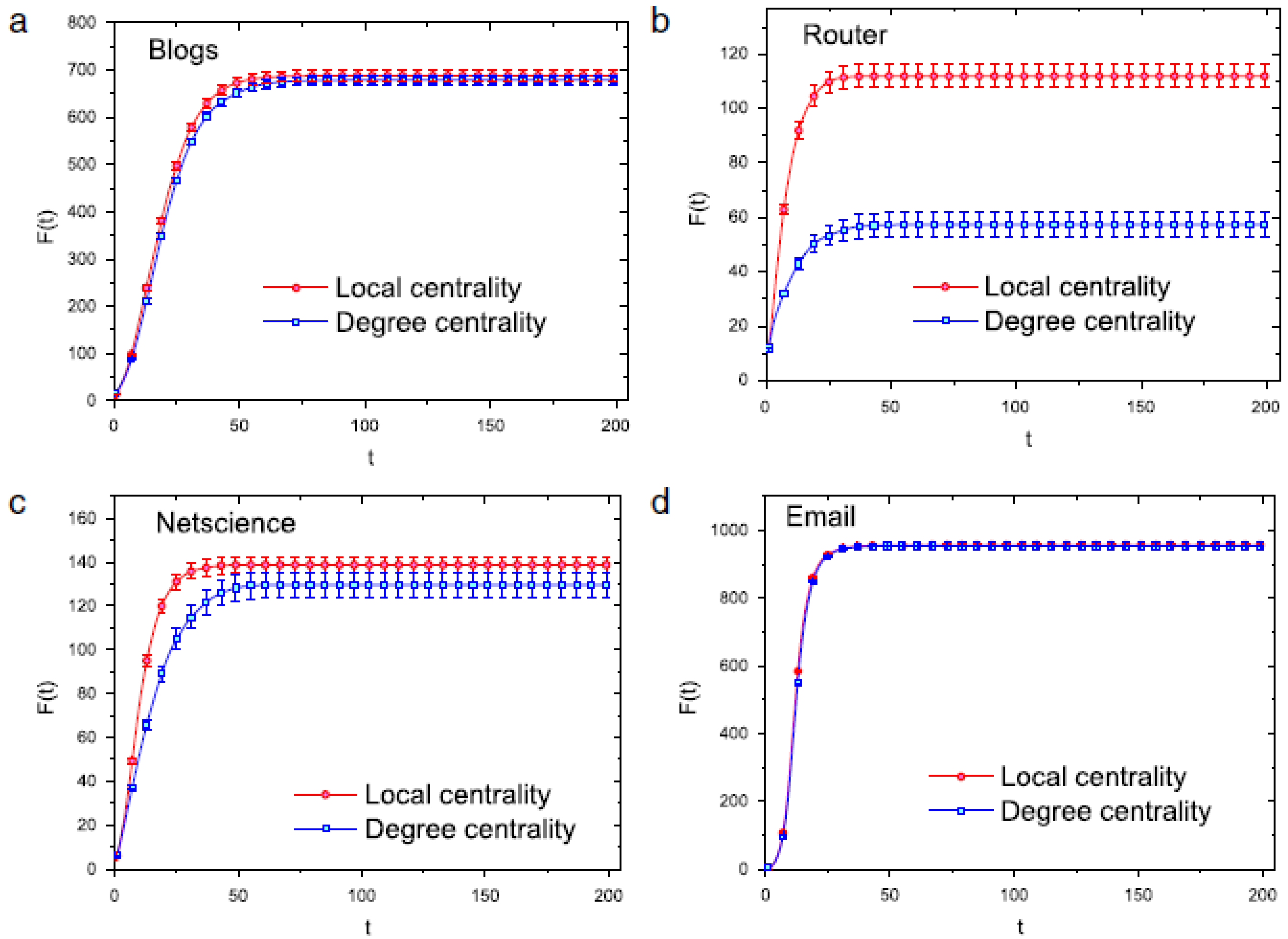


Fig. 4. The cumulative number of infected nodes as a function of time, with the initially infected nodes being those that either appear in the top-10 list by local centrality (circles) or degree centrality (squares), but not appearing in both list. Results are obtained by averaging over 100 implementations.

Back to results

D

ANNEX - RELATIONS BETWEEN LOCAL CENTRALITY AND DEGREE, BETWEENNESS AND CLOSENESS

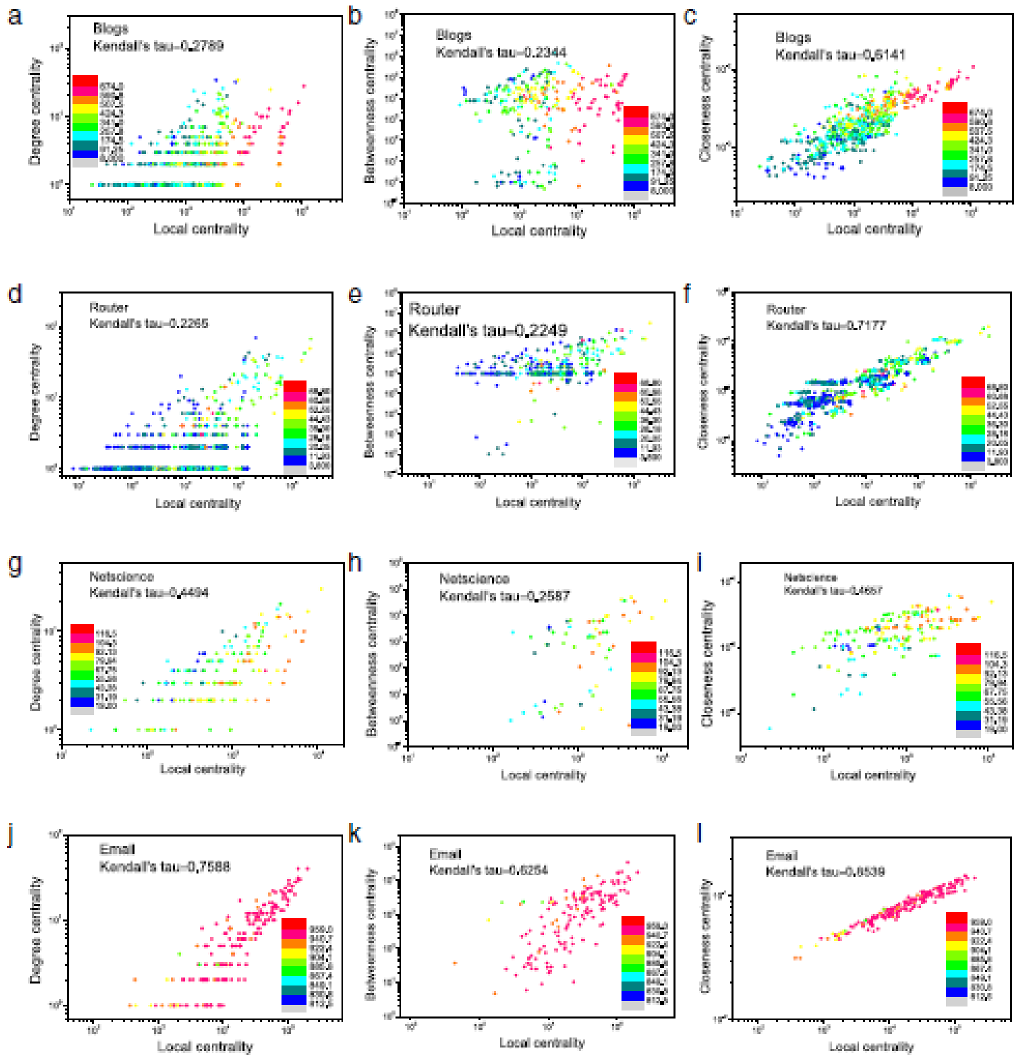


Fig. 6. The relations between local centrality and degree, betweenness and closeness centralities on four example networks. Each data point denotes a node, and its color represent the $F(t)$ value ($t = 10$) of this node. The values are obtained by averaged over 100 independent runs.

Back to results

ANNEX – TABLE (III) TOP-10 RANKED NODES BY LOCAL CENTRALITY AND (IV) MEAN VALUE OF THE TOP NODES

Table 3
The top-10 ranked nodes by local centrality (L) and their corresponding ranks by degree (D), closeness (C) and betweenness (B) centralities. $F(t_c)$ is obtained by averaging over 100 implementations.

Blogs					Router				
L	D	C	B	$F(t_c)$	L	D	C	B	$F(t_c)$
1	1	3	3	525.18	1	5	16	31	7.34
2	15	11	45	671.01	2	9	2	2	8.45
3	3	7	16	520.92	3	7	7	4	7.93
4	37	15	66	652.71	4	6	19	36	7.74
5	16	16	50	679.46	5	11	1	1	8.24
6	87	57	741	619.74	6	15	4	5	12.62
7	96	63	832	660.57	7	20	9	24	8.65
8	88	22	140	610.62	8	16	20	52	8.73
9	109	56	184	653.00	9	37	8	20	7.95
10	135	78	707	624.29	10	31	21	47	12.48
NetScience					Email				
L	D	C	B	$F(t_c)$	L	D	C	B	$F(t_c)$
1	1	19	12	47.08	1	1	3	2	255.70
2	2	7	6	52.38	2	3	4	10	268.18
3	4	77	50	44.16	3	2	1	1	271.70
4	8	81	21	43.30	4	4	40	22	244.38
5	29	85	101	41.92	5	5	2	3	265.24
6	44	86	135	47.56	6	19	11	61	283.24
7	45	87	136	41.40	7	7	19	15	245.54
8	46	88	137	46.78	8	6	5	8	268.64
9	47	89	138	42.58	9	9	21	16	264.62
10	30	22	15	51.84	10	12	33	36	270.24

Table 4
Mean value of $F(t)$ over top-10 nodes on four centralities.

Network	L	D	C	B
Blogs	621.75	373.08	419.28	361.75
Netscience	45.90	41.75	47.21	45.30
Router	40.76	23.81	48.18	37.95
Email	264.75	261.91	267.88	262.88

Back to results

Carles Sans Fuentes
732A61 Data Mining - Clustering and
Association Analysis / Linköpings University
carsa564@student.liu.se

http://www.ida.liu.se/~732A61/timetable/index.en.shtml