BIG DATA ANALYTICS

A Social Network Approach

Andry Alamsyah

Social Computing and Big Data Research Group

Fakultas Ekonomi dan Bisnis

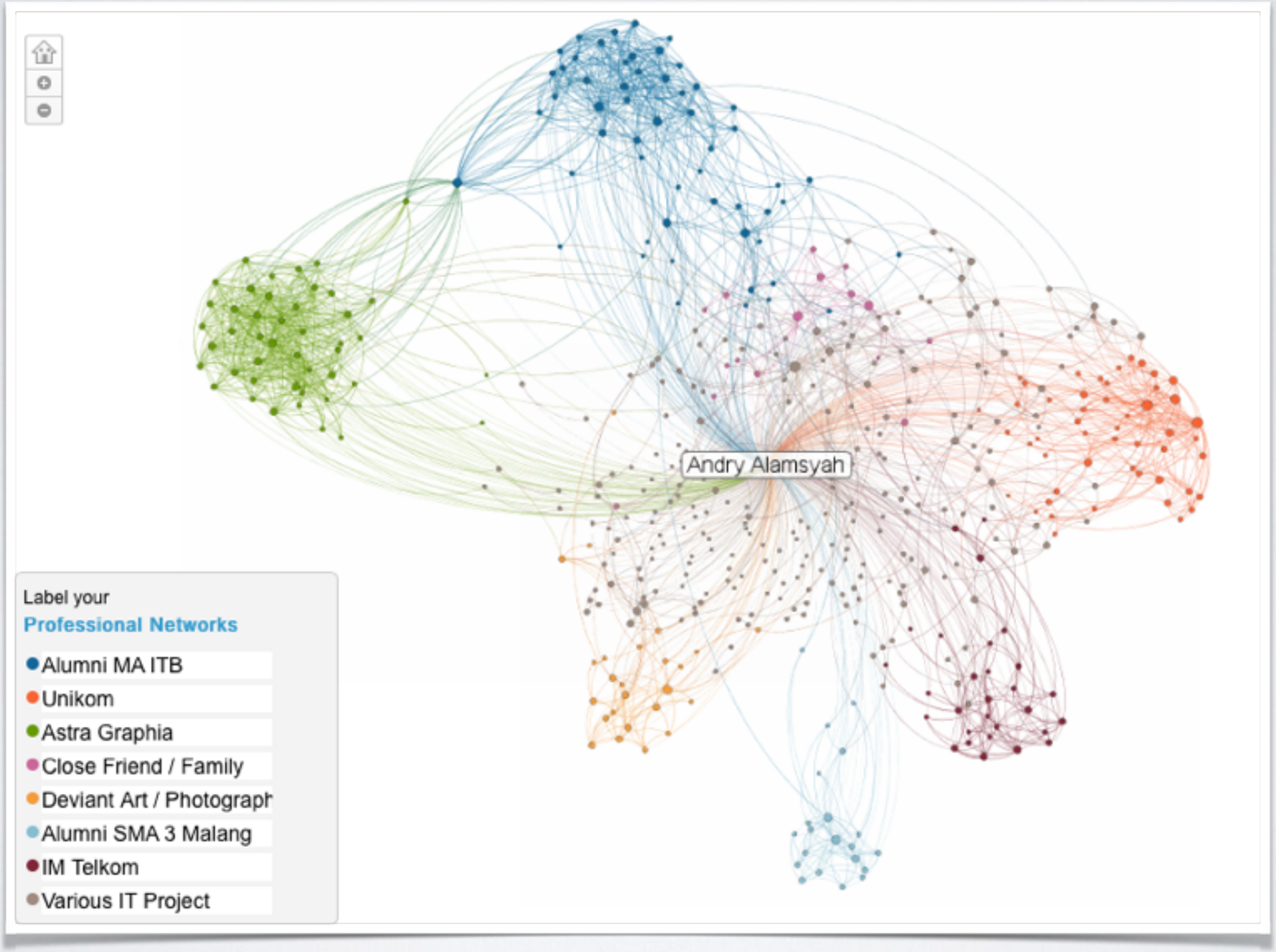
Andry Alamsyah

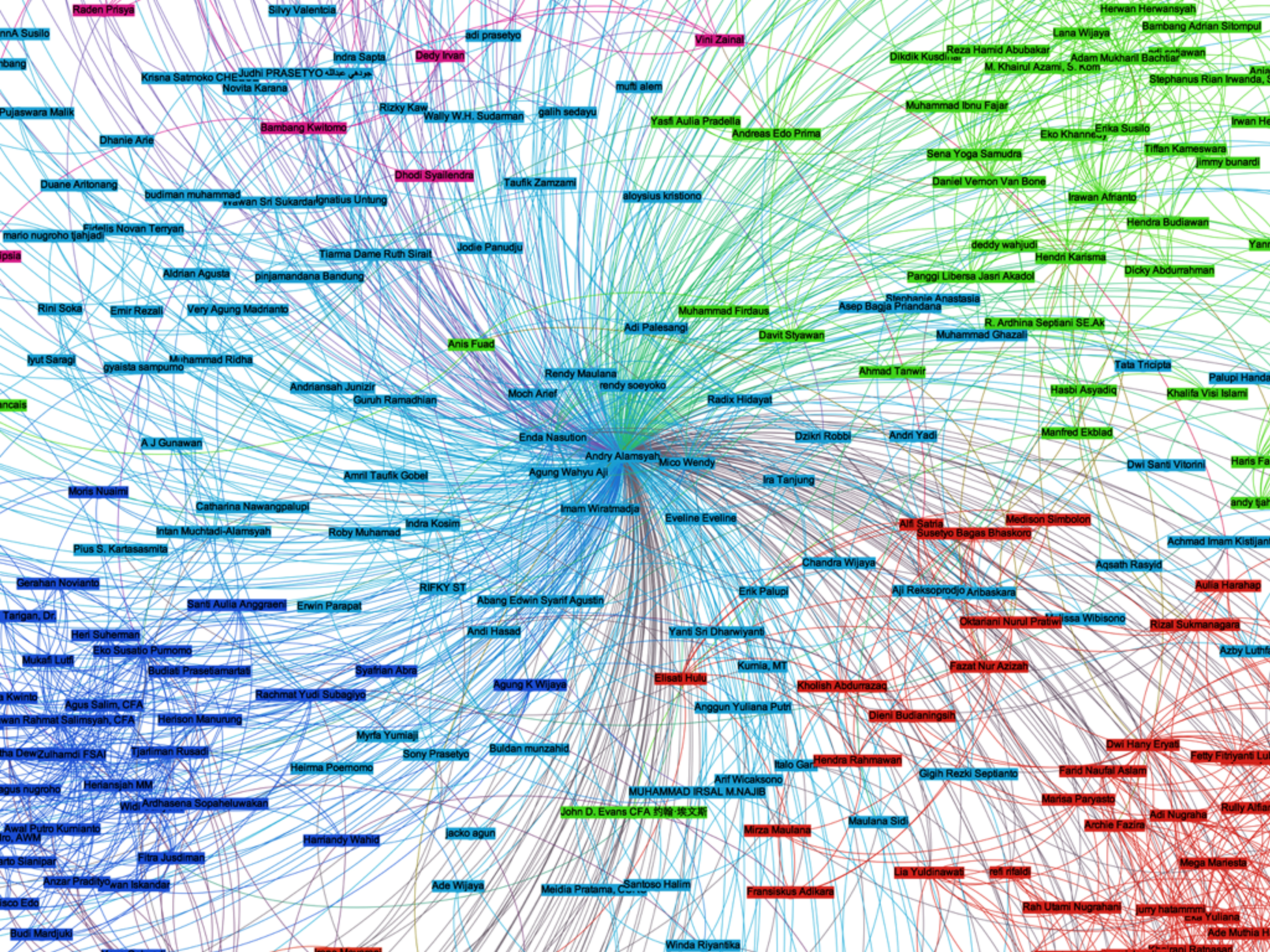
Fakultas Ekonomi dan Bisnis
Universitas Telkom

Research Field :
Social Network, Complex Network / Network Science, Social Computing, Data Analytics, Data Mining, Big Data, Graph Theory, Content Business, Data Business, ICT Business

andry.alamsyah@gmail.com
andrya.staff.telkomuniversity.ac.id
telkomuniversity.academia.edu/andryalamsyah
researchgate.net/profile/Andry_Alamsyah
linkedin.com/andry.alamsyah
twitter.com/andrybrew

WHO AM I ?

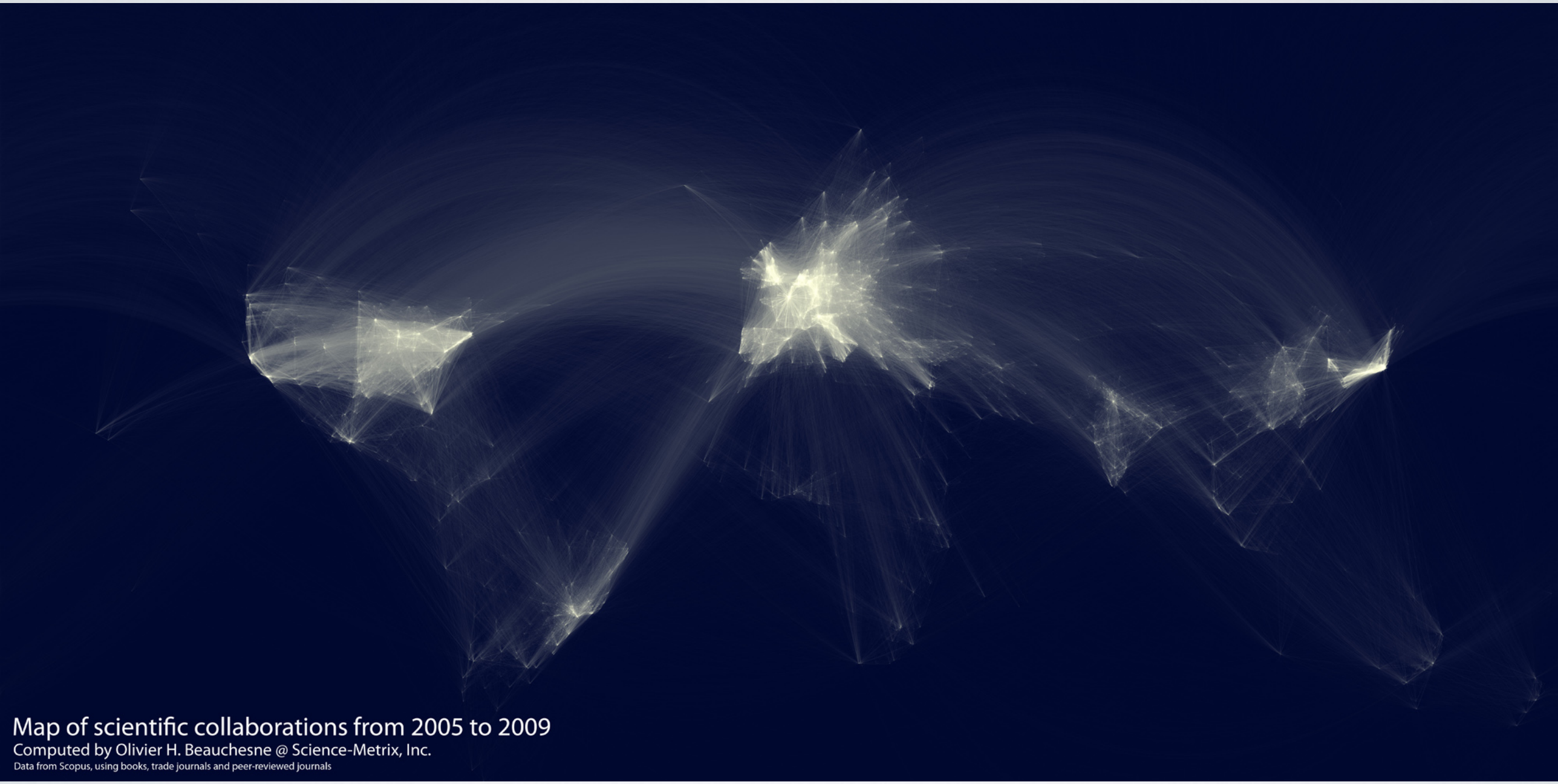

Andry Alamsyah
Label your
Professional Networks
Alumni MA ITB
Unikom
Astra Graphia
Close Friend / Family
Deviant Art / Photograph
Alumni SMA 3 Malang
IM Telkom
Various IT Project


Raden Prisya
Silvy Valentcia
adi prasetyo
Vini Zainal
Herwan Herwansyah
Bambang Adrian Sitompul
Lana Wijaya
Reza Hamid Abubakar
Adam Mukharil Bachtiar
M. Khairul Azami, S. Kom
Stephanus Rian Irwanda, S
Indra Sapta
Dedy Irvan
Judhi PRASETYO
Novita Karana
mufti alem
Muhammad Ibnu Fajar
Rizky Kaw
Wally W.H. Sudarman
galih sedayu
Yasfi Aulia Pradella
Andreas Edo Prima
Erika Susilo
Eko Khannedy
Tiffan Kameswara
Jimmy bunardi
Pujaswara Malik
Bambang Kwitomo
Dhanie Arie
Sena Yoga Samudra
Dhodi Syailendra
Taufik Zamzami
Daniel Vernon Van Bone
Duane Aritonang
budiman muhammad
Ignatius Untung
aloysius kristiono
Irawan Afrianto
Hendra Budiawan
mario nugroho tjahjadi
Jodie Panudju
deddy wahjudi
Hendri Karisma
Tiarma Dame Ruth Sirait
Aldrian Agusta
pinjamandana Bandung
Dicky Abdurrahman
Panggi Libersa Jasri Akadol
Rini Soka
Emir Rezali
Very Agung Madrianto
Muhammad Firdaus
Asep Bagja Priandana
R. Ardhina Septiani SE.Ak
Adi Palesangi
Davit Styawan
Muhammad Ghazali
Anis Fuad
Iyut Saragi
Muhammad Ridha
gyaista sampurno
Ahmad Tanwir
Tata Tricipta
Rendy Maulana
rendy soeyoko
Andriansah Junizir
Moch Arief
Hasbi Asyadiq
Khalifa Visi Islami
Guruh Ramadhian
Radix Hidayat
A J Gunawan
Enda Nasution
Dzikri Robbi
Andri Yadi
Manfred Ekblad
Andry Alamsyah
Mico Wendy
Dwi Santi Vitorini
Amril Taufik Gobel
Agung Wahyu Aji
Ira Tanjung
Moris Nuaimi
Catharina Nawangpalupi
Imam Wiratmadja
Eveline Eveline
Alfi Satria
Medison Simbolon
Susetyo Bagas Bhaskoro
Indra Kosim
Intan Muchtadi-Alamsyah
Roby Muhamad
Achmad Imam Kistijant
Pius S. Kartasasmita
Chandra Wijaya
Aqsath Rasyid
Gerahan Novianto
RIFKY ST
Erik Palupi
Aji Reksoprodjo Aribaskara
Aulia Harahap
Santi Aulia Anggraeni
Erwin Parapat
Abang Edwin Syarif Agustin
Tarigan, Dr.
Oktariani Nurul Pratiwi
Rizal Sukmanagara
Heri Suherman
Andi Hasad
Yanti Sri Dharwiyanti
Eko Susatio Purnomo
Mukafi Lutfi
Kurnia, MT
Budiati Prasetiamartati
Syafrian Abra
Fazat Nur Azizah
Elisati Hulu
Agung K Wijaya
Kholish Abdurrazaq
Rachmat Yudi Subagiyo
Agus Salim, CFA
Anggun Yuliana Putri
Herison Manurung
Dieni Budianingsih
Myrfa Yumiaji
Dwi Hany Eryati
Tjarliman Rusadi
Buldan munzahid
Sony Prasetyo
Hendra Rahmawan
Heirma Poernomo
Gigih Rezki Septianto
Farid Naufal Aslam
Herlansjah MM
Arif Wicaksono
MUHAMMAD IRSAL M.NAJIB
Marisa Paryasto
Ardhasena Sopaheluwakan
John D. Evans CFA 约翰·埃文斯
Adi Nugraha
Archie Fazira
Awal Putro Kurnianto
Jacko agun
Mirza Maulana
Maulana Sidi
Harriandy Wahid
Fitra Jusdiman
Mega Mariesta
Lia Yuldinawati
refi rifaldi
Anzar Pradityo
Santoso Halim
Ade Wijaya
Fransiskus Adikara
Rah Utami Nugrahani
Budi Mardjuki
Winda Riyantika

LARGE SCALE DATA

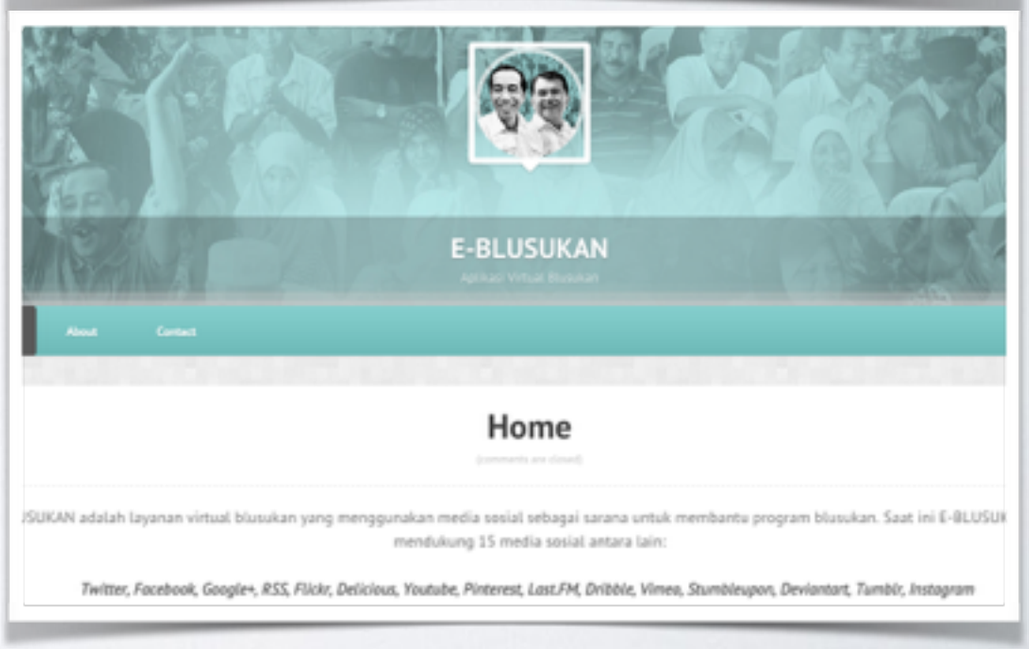
LARGE SCALE DATA

LARGE SCALE DATA



STORY / PHENOMENON

- BIG DATA leads to Social Computing *(Quantification of Individual / Social Behaviour)*
- Social Network Data / Conversation are widely available
- Social Network voices represent public voice become ***'Big'*** concern (references)
- The Need of Real-Time Analytic *(OLAP)*
- The Need of Powerful Metric for Social Network / Big Data



STORY / PHENOMENON

- There are many aspect of Big Data research, but too little resource, too little talent

- Same business objective, but increase effectiveness on top of current services

- Problem with Legacy Methodology approach using Questionnaire/ Interviews/Surveys *(ok with small scale data , expensive and took longer time for large scale data, accuracy issues)*

INDUSTRY EFFORTS

METHODS COMPARISON IN SOCIAL SCIENCE

LEGACY	DATA ANALYTICS
Confirmative	Explorative (Predictive)
Small Data Set	Larga Data Set
Small Number of Variable	Large Number of Variable
Deductive (no predictions)	Inductive
Numeric Data	Numeric and Non-Numeric Data
Clean Data	Data Cleaning

source : Data Mining and Statistics :Whats the Connections ? (Jerome Friedman)

BIG DATA STATE OF THE ART

Computation Related

Processing / Computation	Storage	Analytics Tools
• Hadoop • Nvidia CUDA • Twitter Storm • Bulk Synchronous Parallel Processing • GraphLab • Disk-Based Graph Processing	• neo4J • Titan • HDFS	• MLPACK • Mahout

Methodology / Analytics Related

modelling, descriptions, predictions, optimisation and simulation

Big Data - State of the art - Chalmers et al (2013)

BIG DATA ANALYTICS CONSTRUCTOR

Social Network

networks

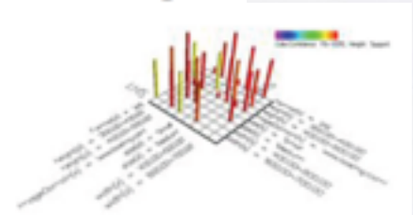
tie-strength

key players

cohesion

Data Mining

Clustering



Association Rules

Classification & Regression

Anomaly Detection

Sentiment Analysis

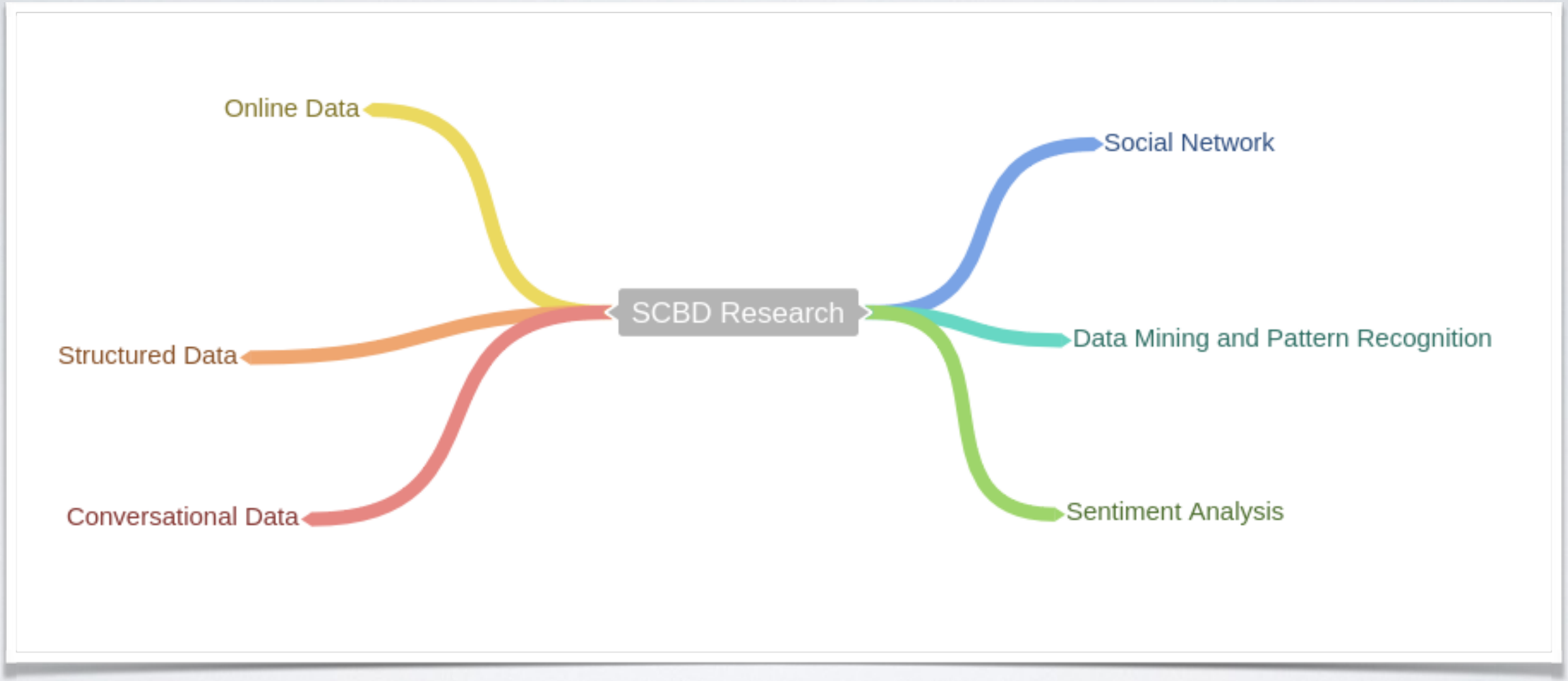
keyword spotting

lexical affinity

statistical methods

concept-level technique

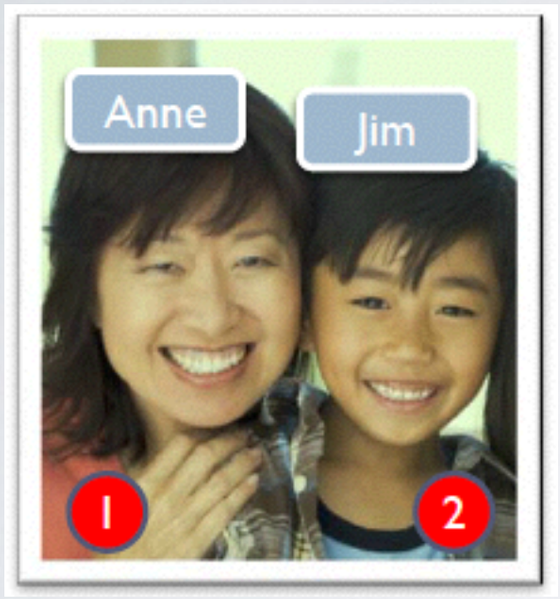
RESEARCH ROADMAP

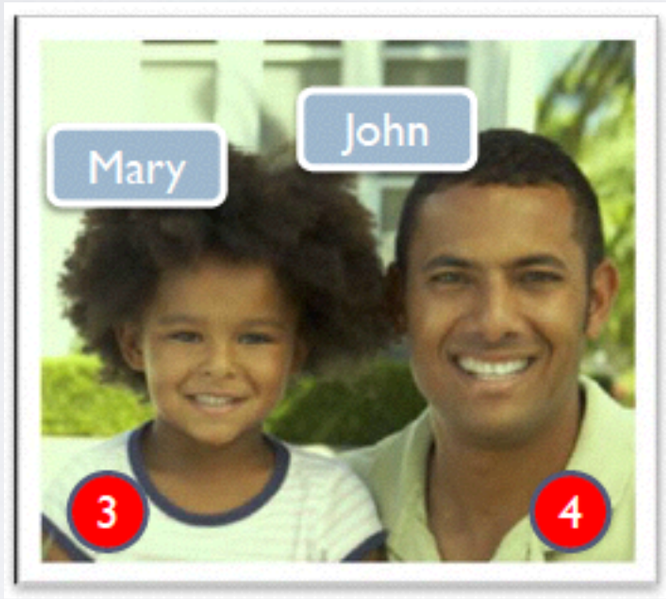


GOAL : descriptions, predictions, optimisation and simulation

area : marketing, communications, knowledge management, operations, finance, etc

SOCIAL NETWORK MODEL





Can we study their interactions as a network ?

Communication

Anne: Jim, tell the Murrays they're invited

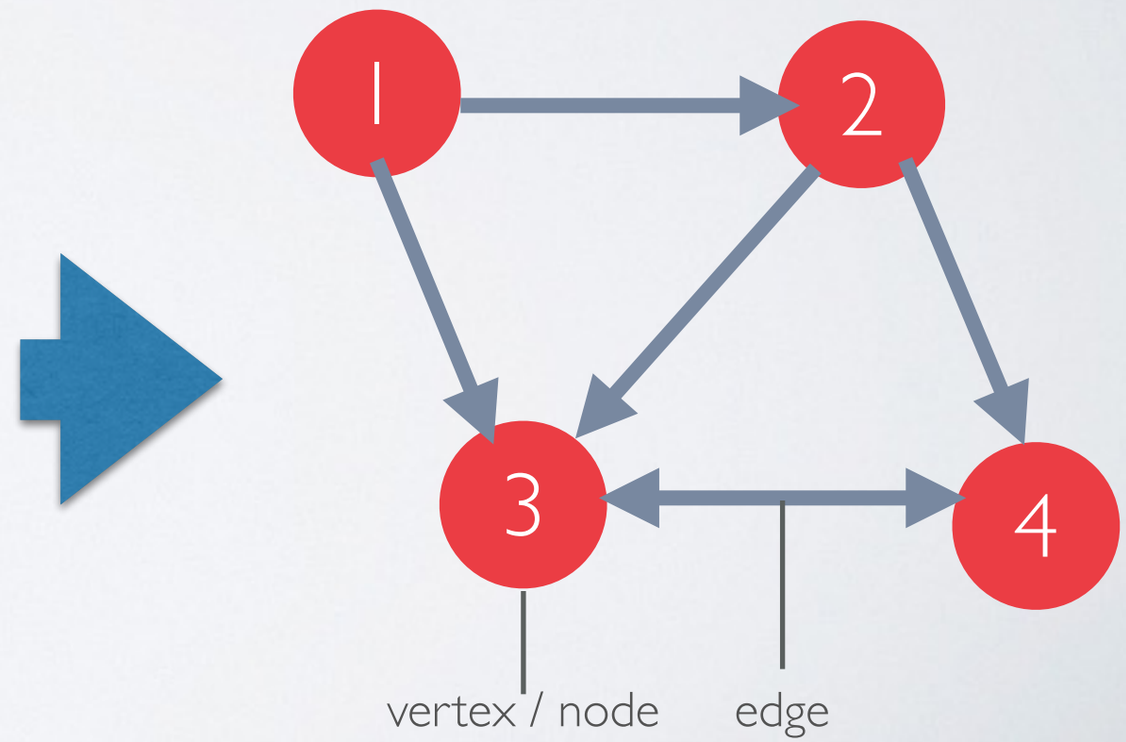
Jim: Mary, you and your dad should come for dinner!

Jim: Mr. Murray, you should both come for dinner

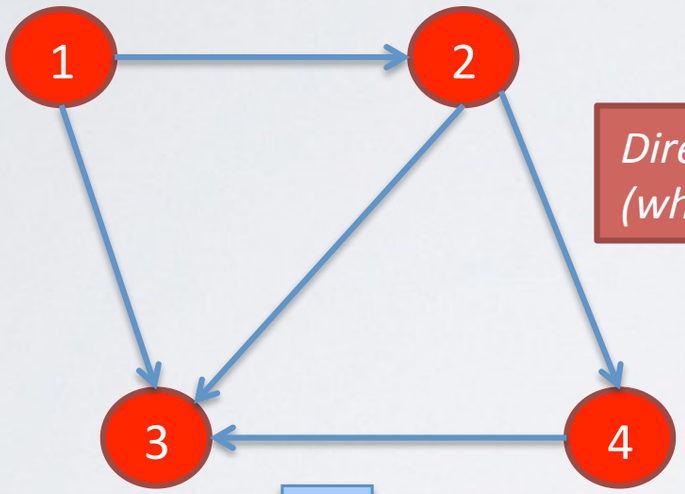
Anne: Mary, did Jim tell you about the dinner? You must come.

Mary: Dad, we are invited for dinner tonight

John: *(to Anne)* Ok, we're going, it's settled!



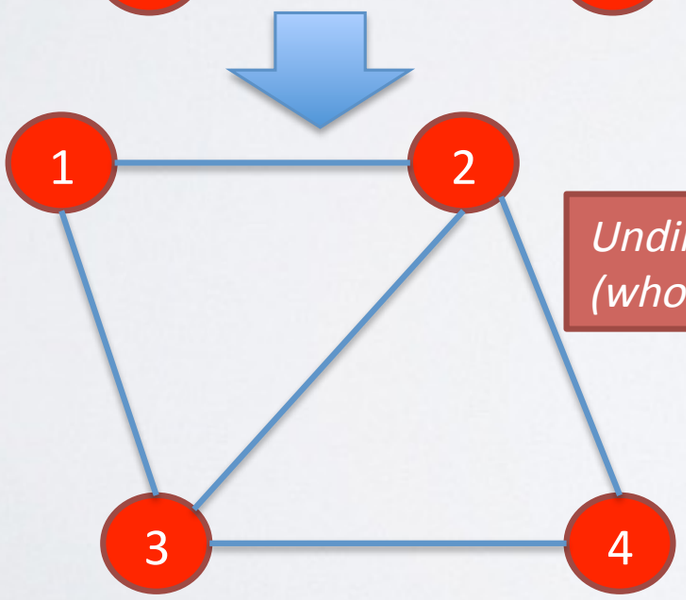
SOCIAL NETWORK MODEL



Directed
(who contact whom)

Edges List

Vertex	Vertex
1	2
1	3
2	3
2	4
4	3

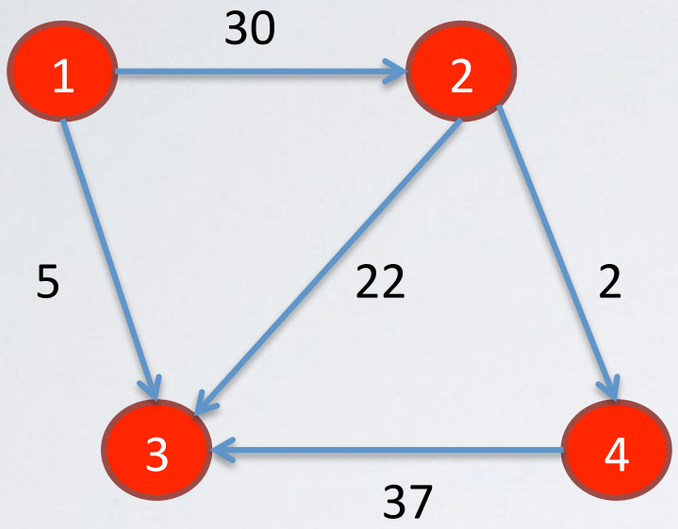


Undirected
(who knows whom)

Adjacency Matrix become symmetric

Vertex	1	2	3	4
1	-	1	1	0
2	1	-	1	1
3	1	1	-	0
4	0	1	0	-

TIE STRENGTH



Weight could be
- Frequency of interactions in period of observation
- Number of items exchanged in period
- Individual perceptions of strength of relationship
- Cost of communications or exchange, e.g. distance

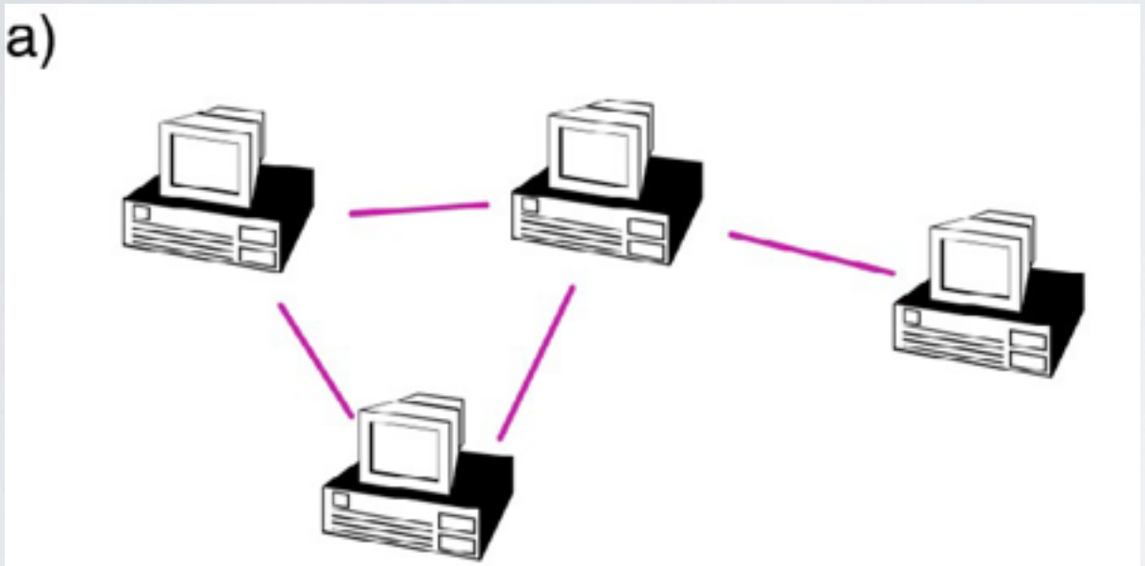
Edges List

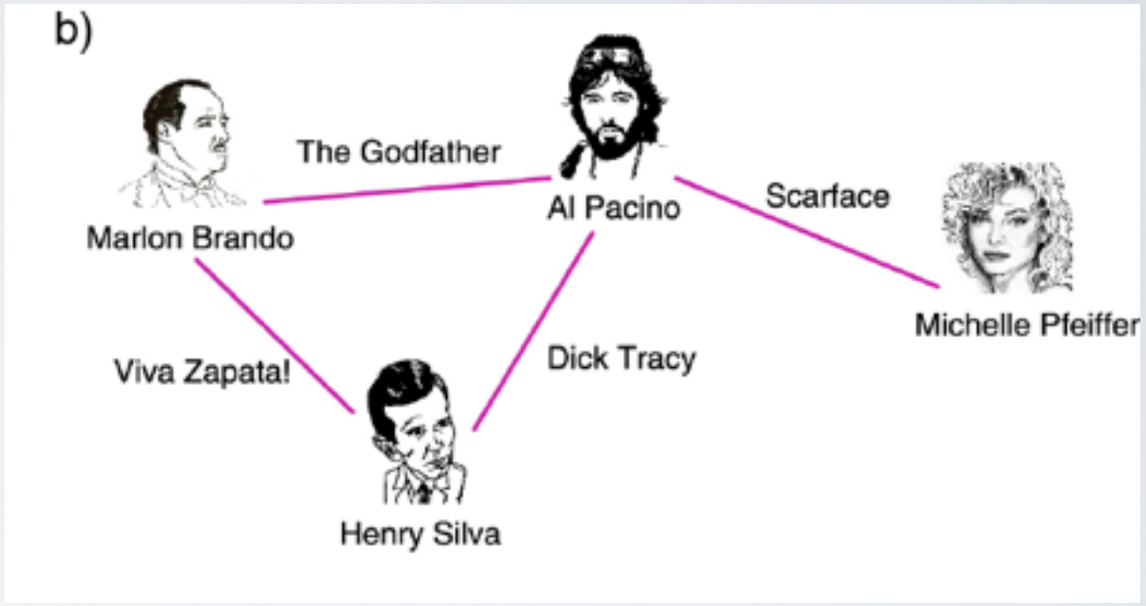
Vertex	Vertex	Weight
1	2	30
1	3	5
2	3	22
2	4	2
4	3	27

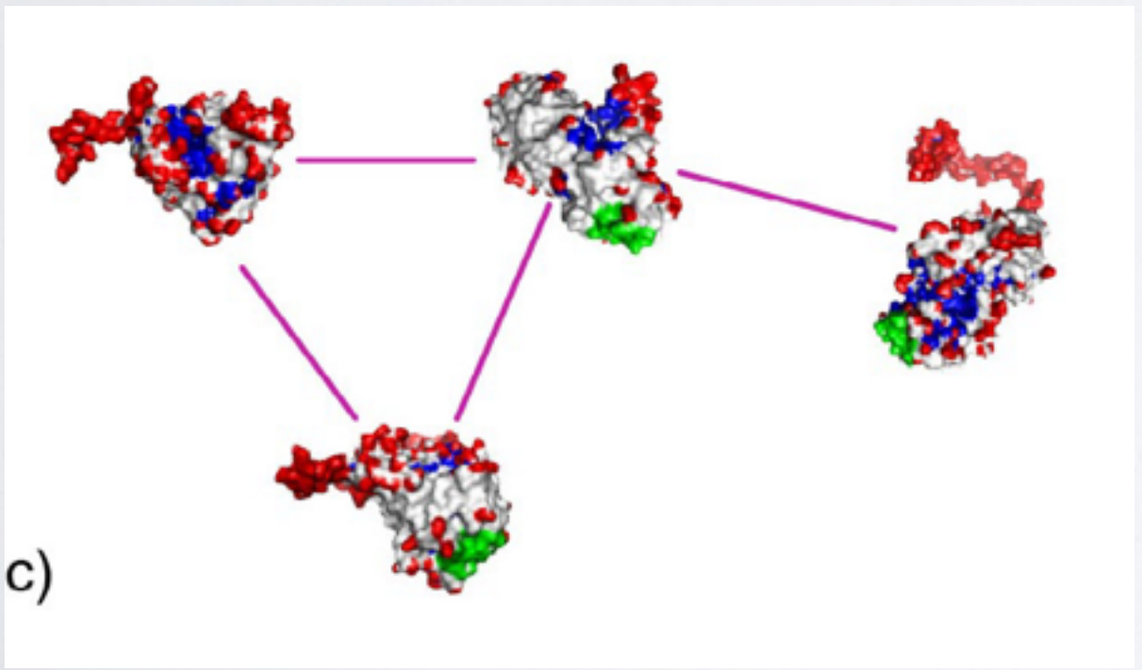
Adjacency Matrix (weight)

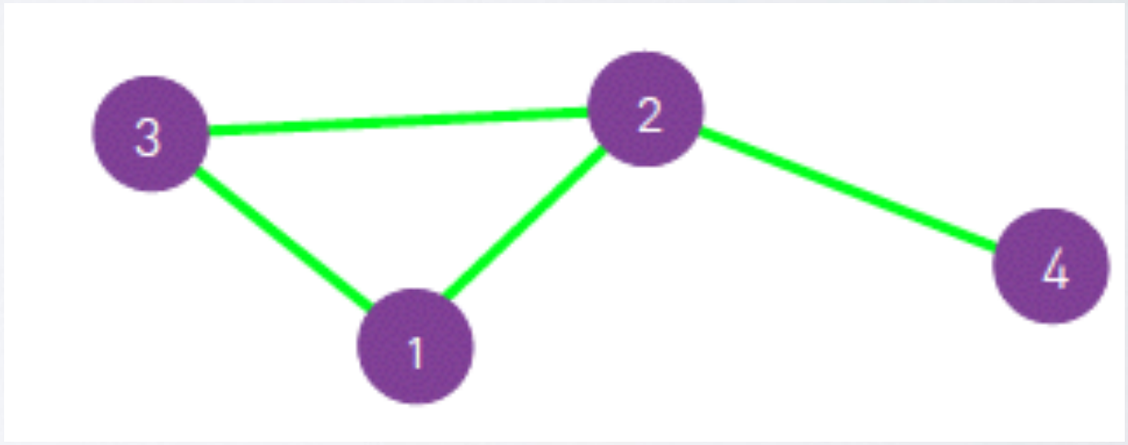
Vertex	1	2	3	4
1	-	30	5	0
2	30	-	22	2
3	5	22	-	37
4	0	2	37	-

NETWORK MODEL EXAMPLE



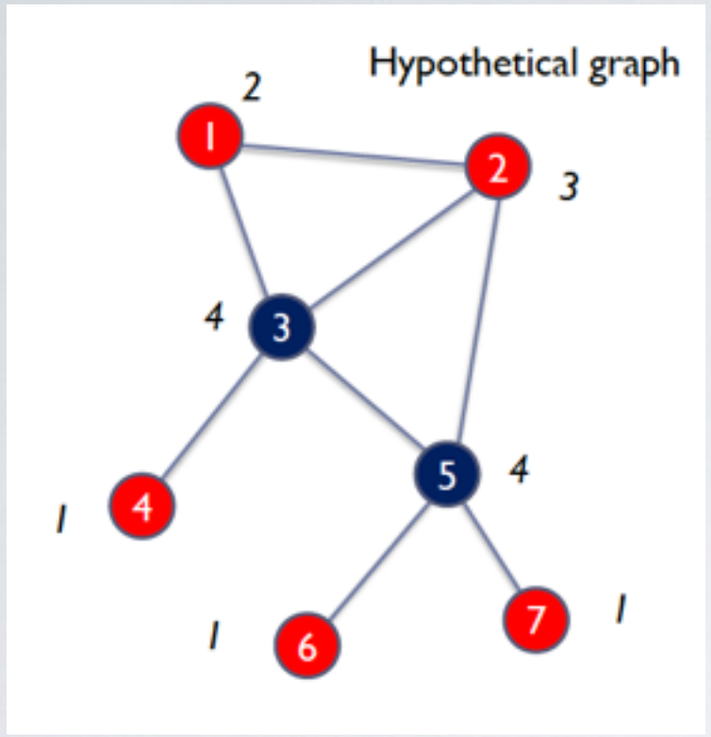






Different Network, Same Graph

METRIK CENTRALITY



degree centrality

shortest path

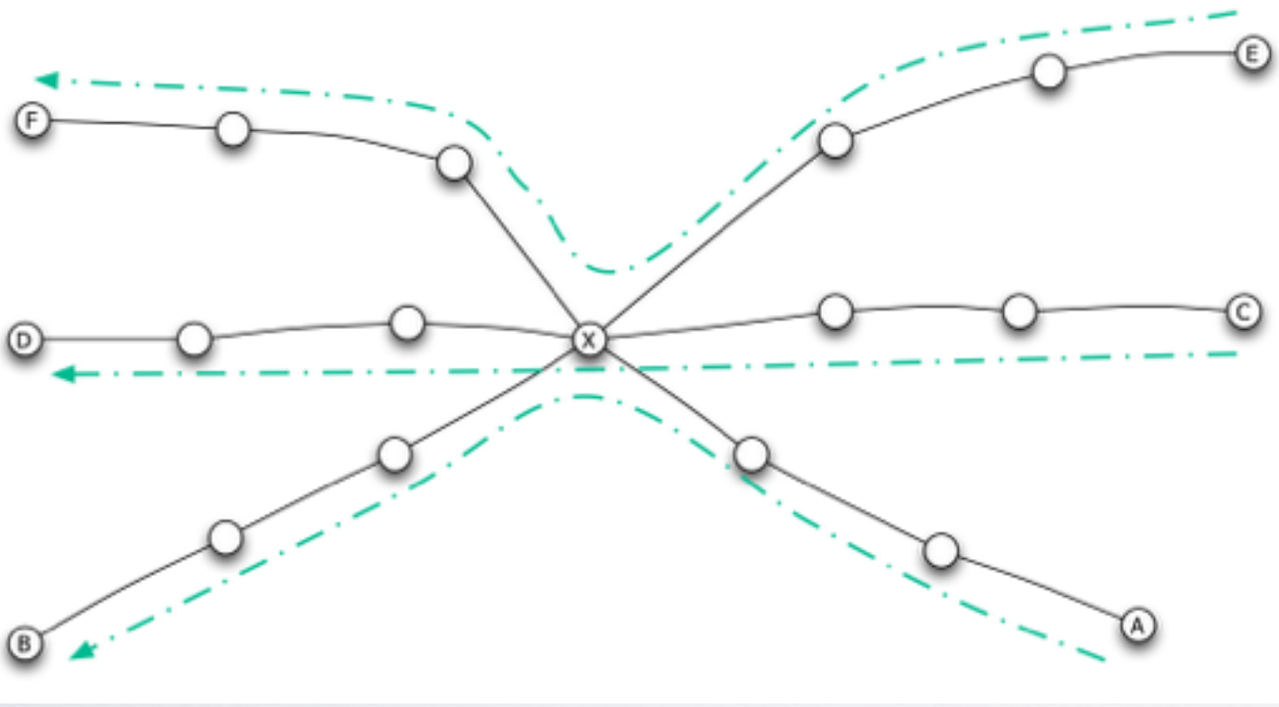
betweenness centrality

closeness centrality

eigenvector centrality

set of key players

METRIK CENTRALITY

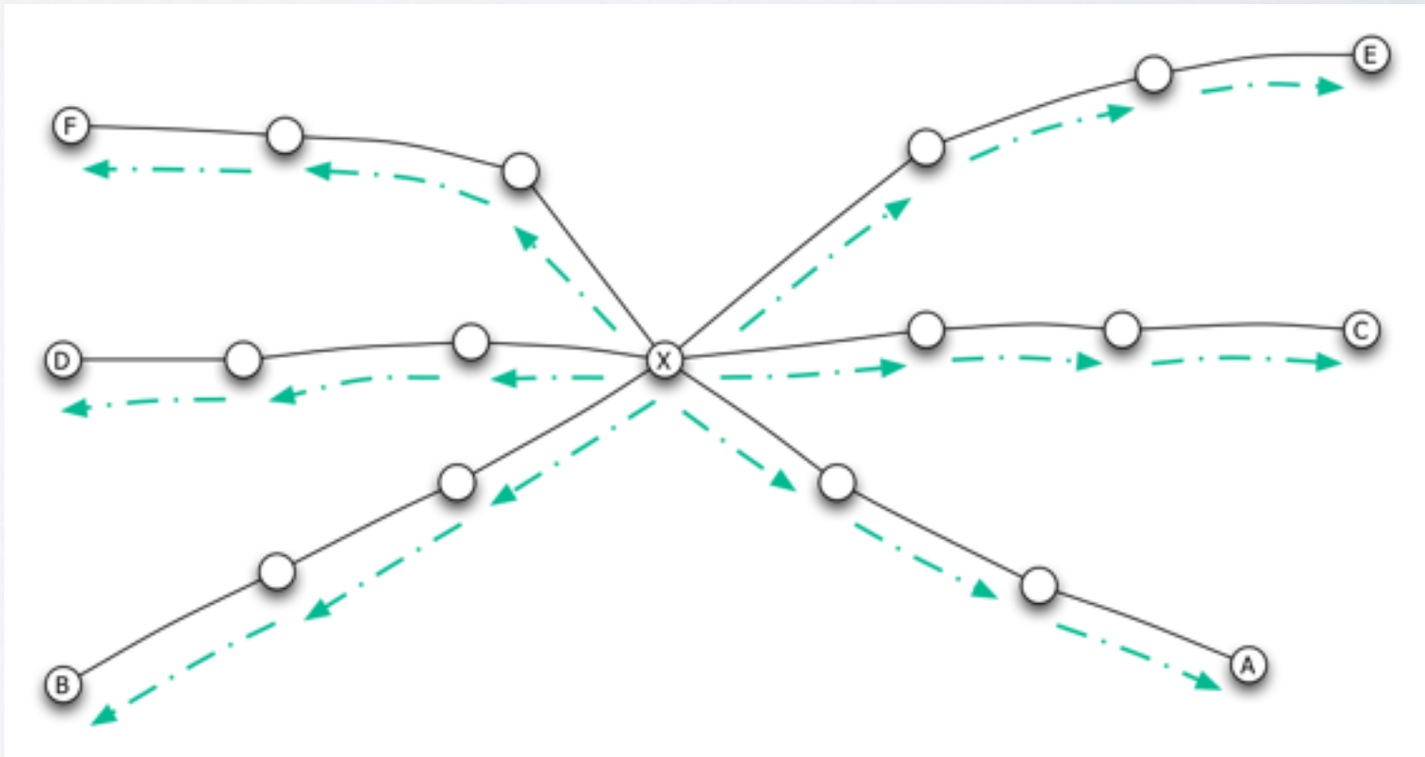


betweenness centrality

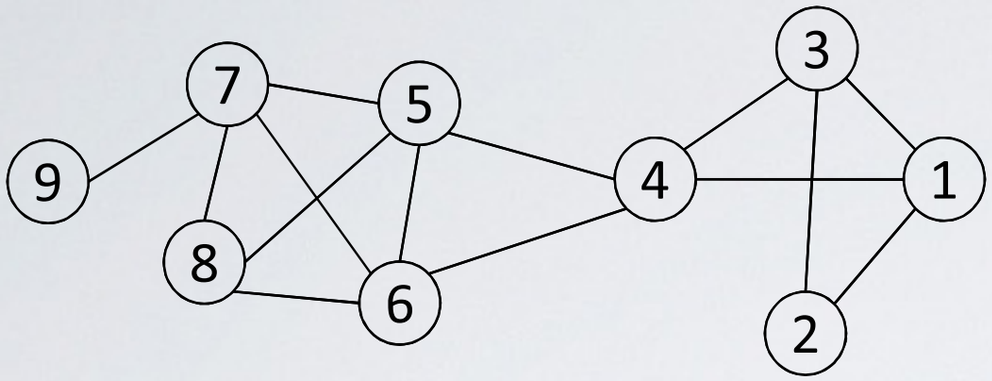
banyaknya **jalur terpendek** antar pasangan semua titik di jaringan, yang melewati satu titik yang diukur

closeness centrality

jarak titik yang diukur terhadap semua titik yang ada dalam jaringan



FORMULASI CENTRALITY



kompleksitas waktu perhitungan metrik mencapai* ***O(n³)* dan kompleksitas ruang sebesar ***O(n²)***

Betweenness Centrality

$$C_B(i) = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

$$\sigma_{st}(4) / \sigma_{st}$$

	s=1	s=2	s=3
t=5	1/1	2/2	1/1
t=6	1/1	2/2	1/1
t=7	2/2	4/4	2/2
t=8	2/2	4/4	2/2
t=9	2/2	4/4	2/2

$$C_B(4) = 15$$

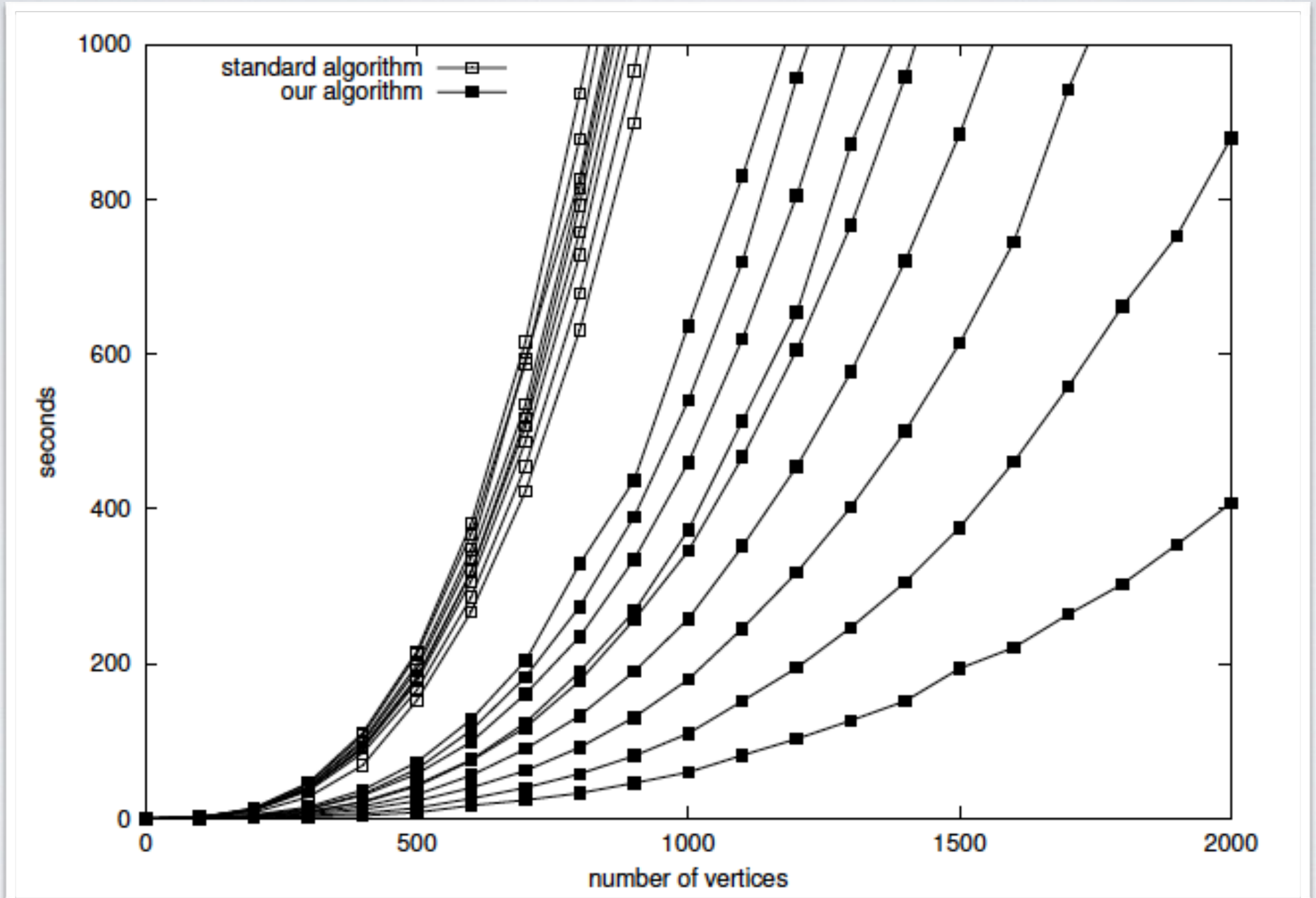
Closenness Centrality

$$C_c(i) = \frac{n-1}{\sum_{j(\neq i)} d_{ij}}$$

	1	2	3	4	5	6	7	8	9
1	0	1	1	1	2	2	3	3	4
2	1	0	1	2	3	3	4	4	5
3	1	1	0	1	2	2	3	3	4
4	1	2	1	0	1	1	2	2	3
5	2	3	2	1	0	1	1	1	2
6	2	3	2	1	1	0	1	1	2
7	3	4	3	2	1	1	0	1	1
8	3	4	3	2	1	1	1	0	2
9	4	5	4	3	2	2	1	2	0

$$C_c(4) = 0{,}62 \qquad C_c(3) = 0{,}47$$

COMPUTATION PROBLEM



paper: A Faster Algorithm for Betweenness Centrality - Brandes (2004)

CENTRALITY INTERPRETATIONS

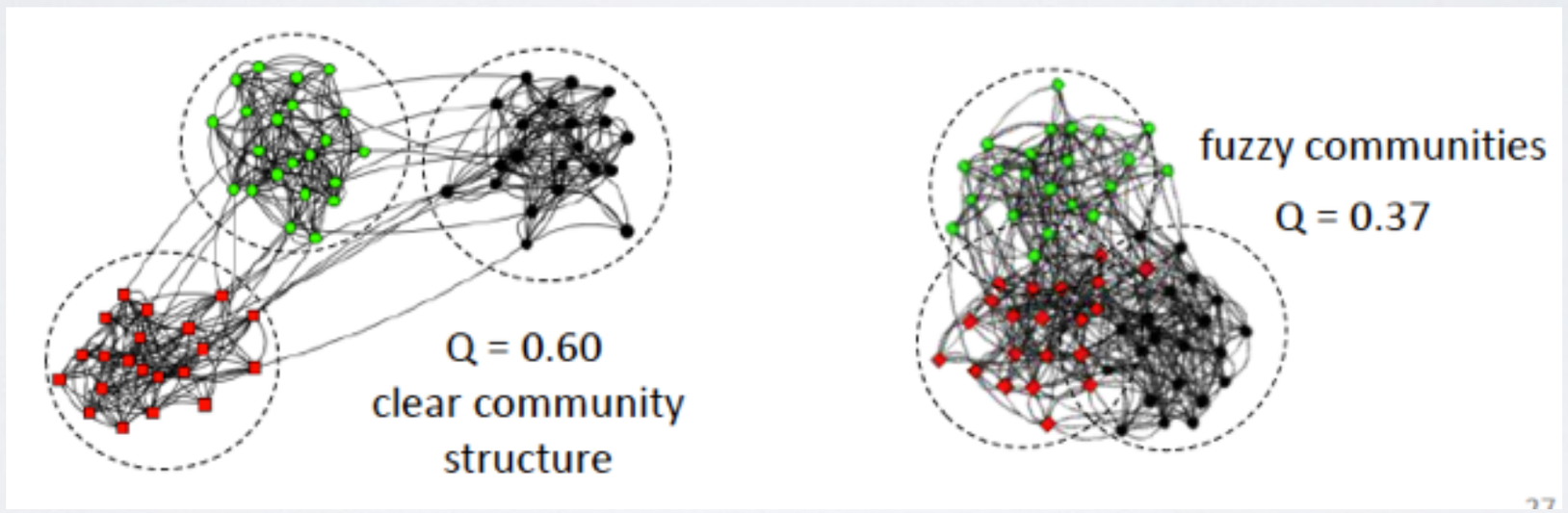
Centrality Measures	Interpretation 1	Interpretation 2
Degree	*how many people can this person reach directly ?*	*in network of investors and start up: how many start up has this company invested to ?*
Betweenness	*how likely is this person to be the most direct route between two people in the network ?*	*in network of knowledge: who is the employee though whom most of the confidential information is likely to flow ?*
Closeness	*how fast can this person reach everyone in the network ?*	*in network of information dissemination: how fast this word-of-mouth issue spread from this person to the rest of the network ?*
Eigenvector	*how well is this person connected to other well-connected people ?*	*in network of paper citations: who is the author that is most cited by other well-cited authors ?*

METRIK MODULARITY

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(C_i, C_j)$$

edges inside the community

expected number of edges if *i,j* places at random



Metrik Modularity

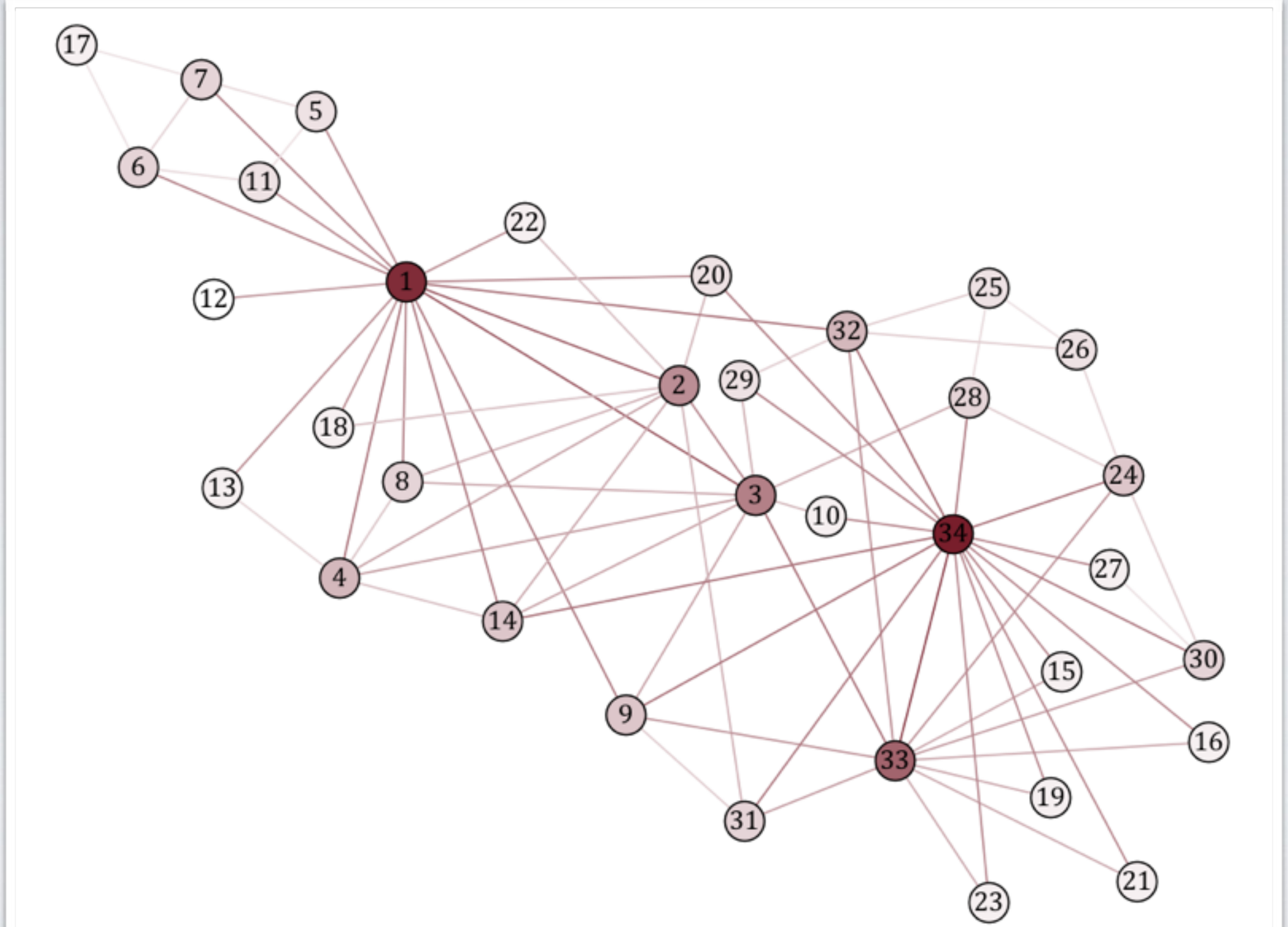
indeks kualitas partisi jaringan menjadi komunitas

Kompleksitas Metrik

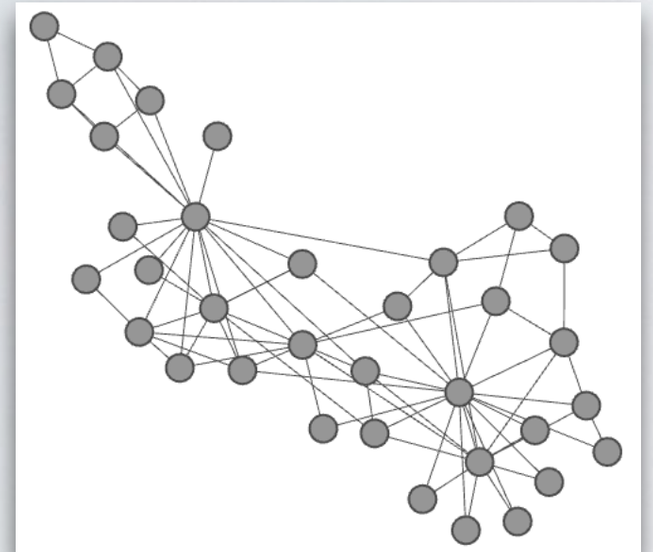
kompleksitas waktu perhitungan

$O(n^3)$

NETWORK MODEL EXAMPLE

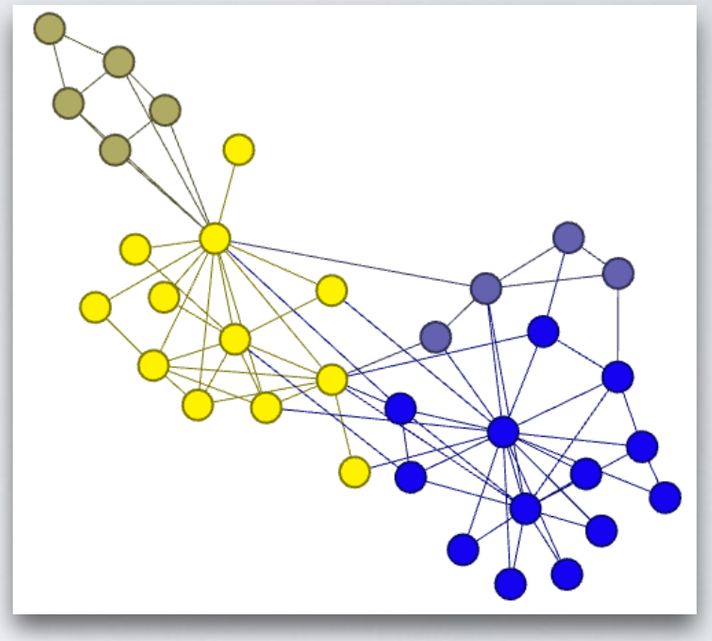


METRIC SIMULATION



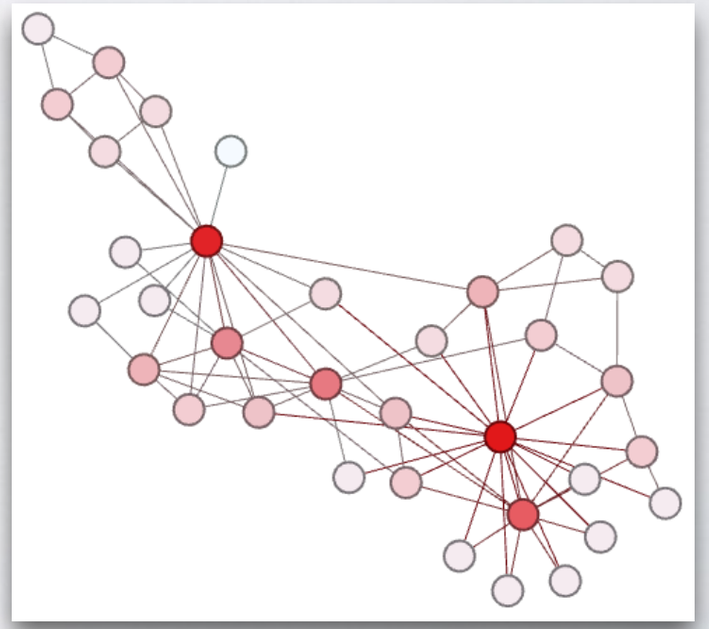
model network

modularity
community detection



community detection result

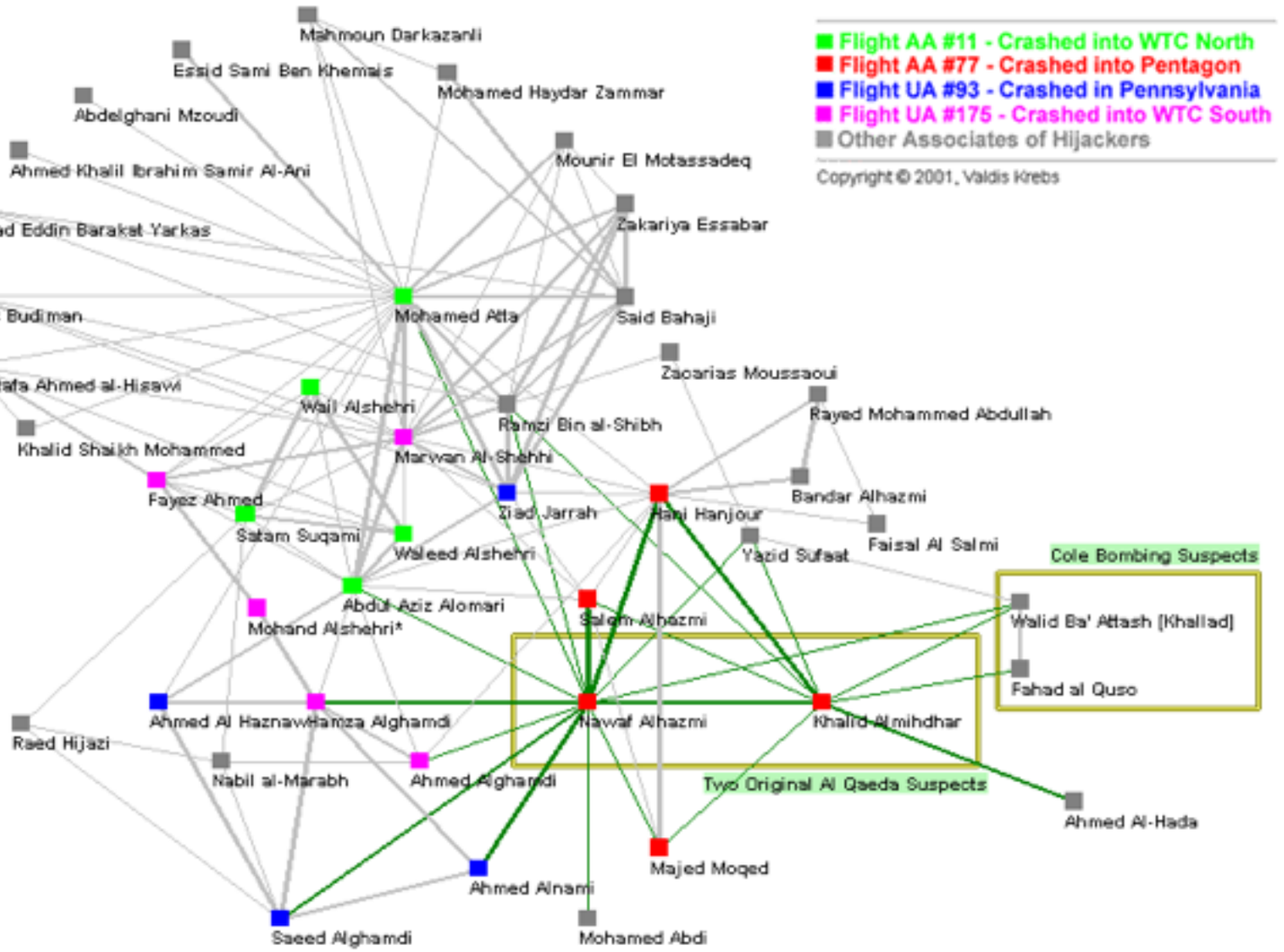
centrality



degree centrality result

**karate club dataset*
34 nodes dan 78 edges

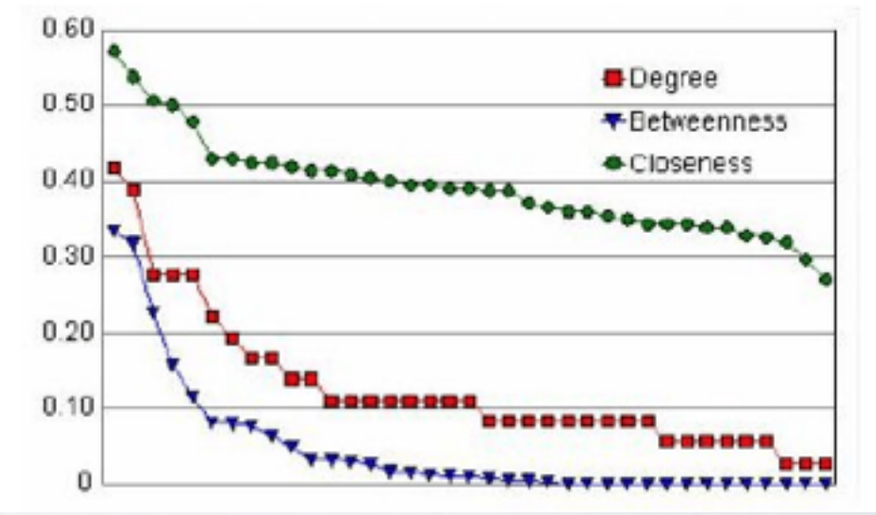
EXAMPLE : FINDING INFLUENCER



Group Size	37
Potential Ties	1332
Actual Ties	170
Density	13%

Geodesics

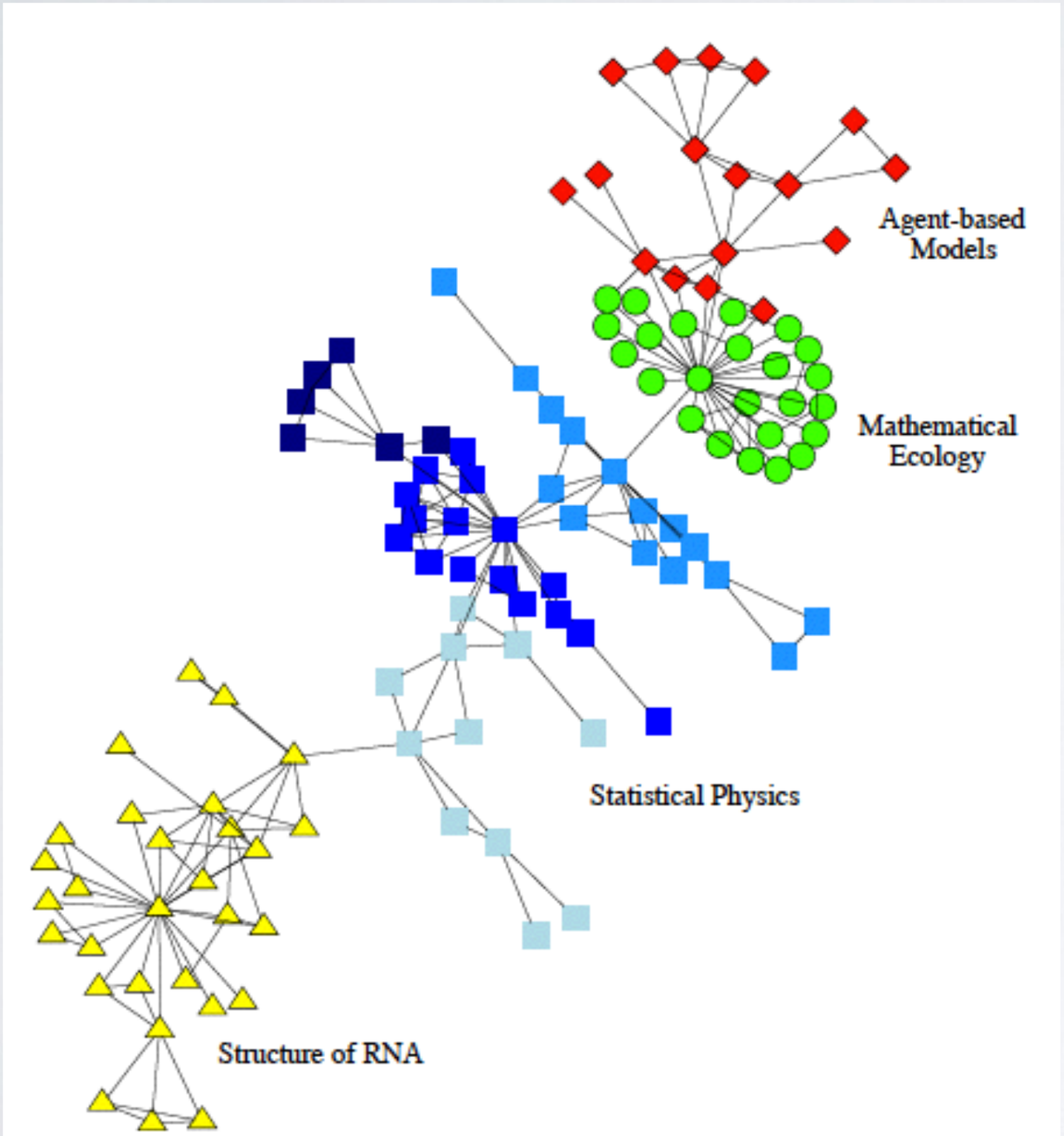
length	#
1	170
2	626
3	982
4	558
5	136
6	0



Degrees		Betweenness		Closeness	
0.417	Mohamed Atta	0.334	Nawaf Alhazmi	0.571	Mohamed Atta
0.389	Marwan Al-Shehhi	0.318	Mohamed Atta	0.537	Nawaf Alhazmi
0.278	Hani Hanjour	0.227	Hani Hanjour	0.507	Hani Hanjour
0.278	Nawaf Alhazmi	0.158	Marwan Al-Shehhi	0.500	Marwan Al-Shehhi
0.278	Ziad Jarrah	0.116	Saeed Alghamdi*	0.480	Ziad Jarrah
0.222	Ramzi Bin al-Shibh	0.081	Hamza Alghamdi	0.429	Mustafa al-Hisawi

paper: Mapping Network of Terrorist Cells - V.E.Krebs (2002)

EXAMPLE : FINDING COMMUNITY

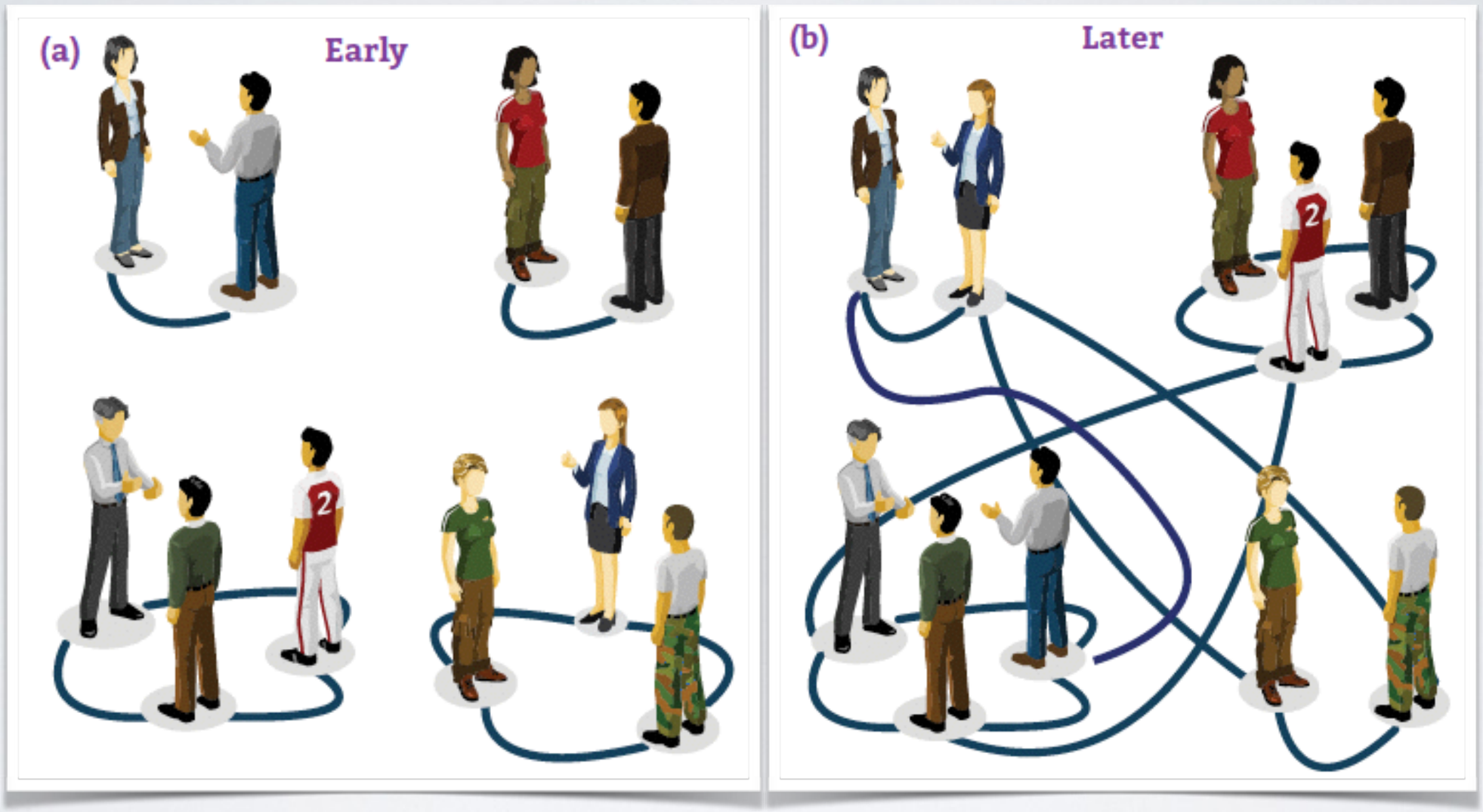


collaboration network of scientist at Santa Fe Institut (Girvan & Newman)
271 scientist (vertices) / 118 nodes from largest component
edge = scientist coauthor one of more publications

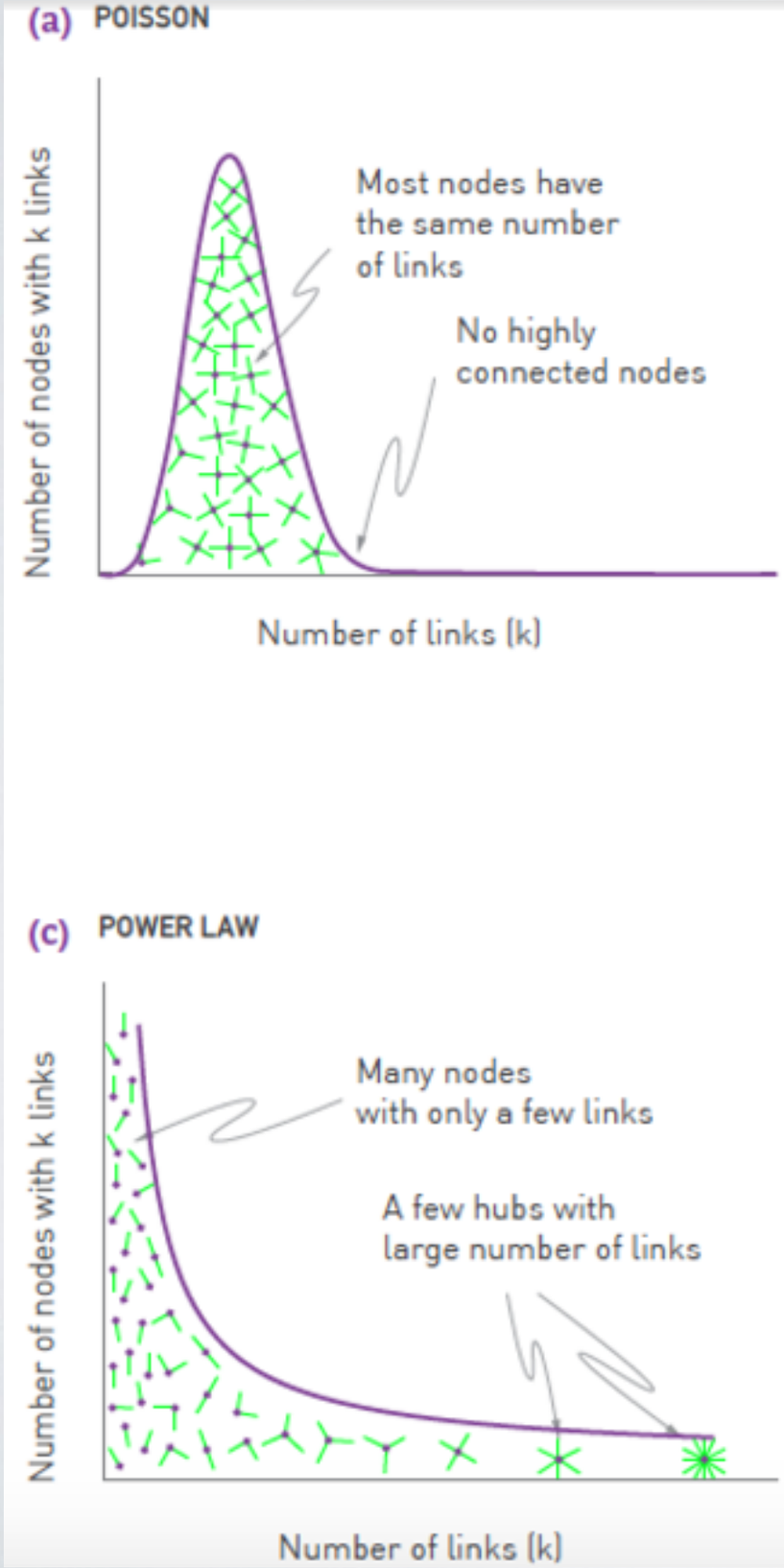
Komunitas : *kumpulan titik titik dimana jumlah hubungan internal antar titik lebih besar dari pada jumlah hubungan dengan titik eksternal*

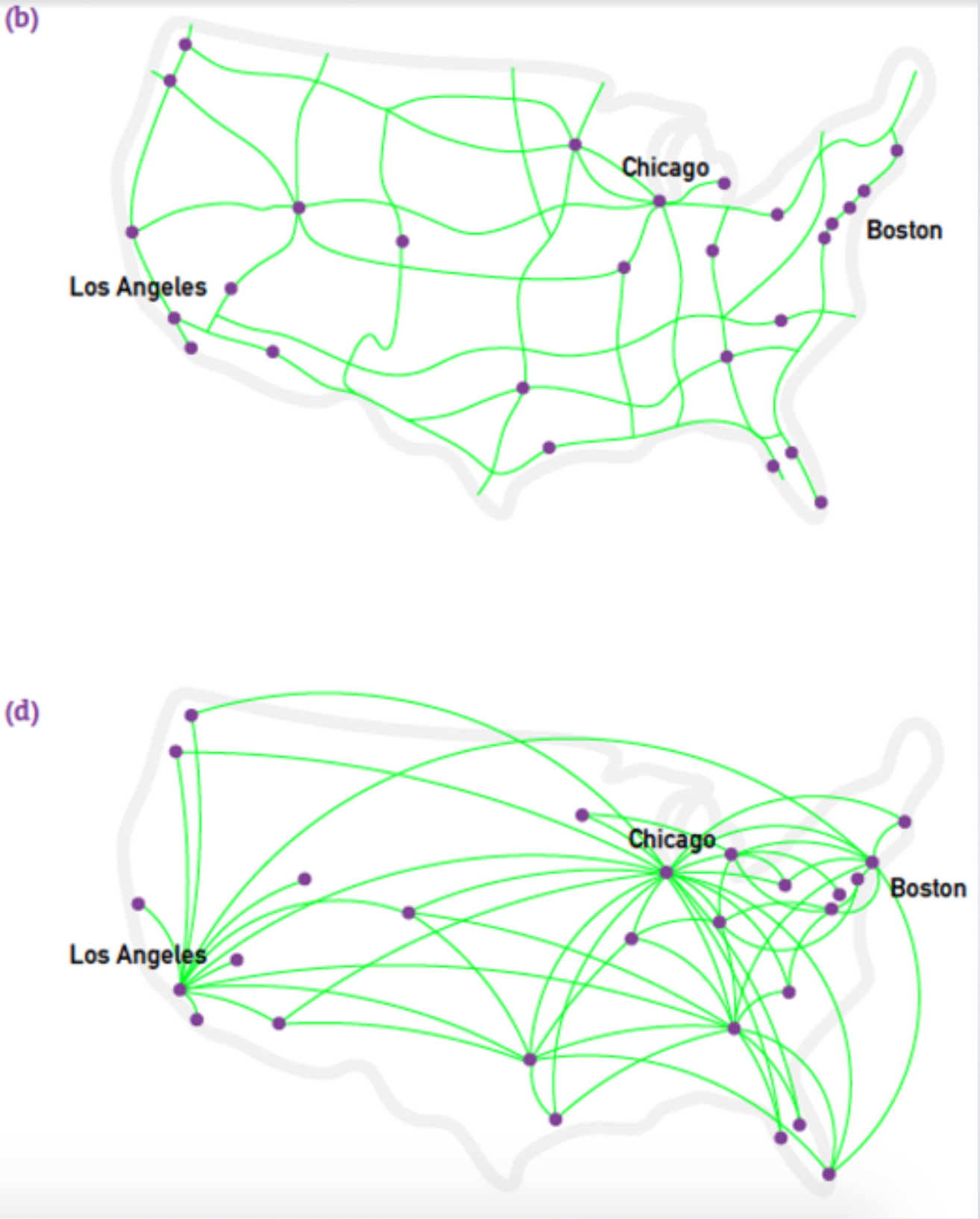
paper: Community Structure in Social and Biological Networks - Girvan-Newman (2002)

SOCIAL NETWORK FORMATIONS

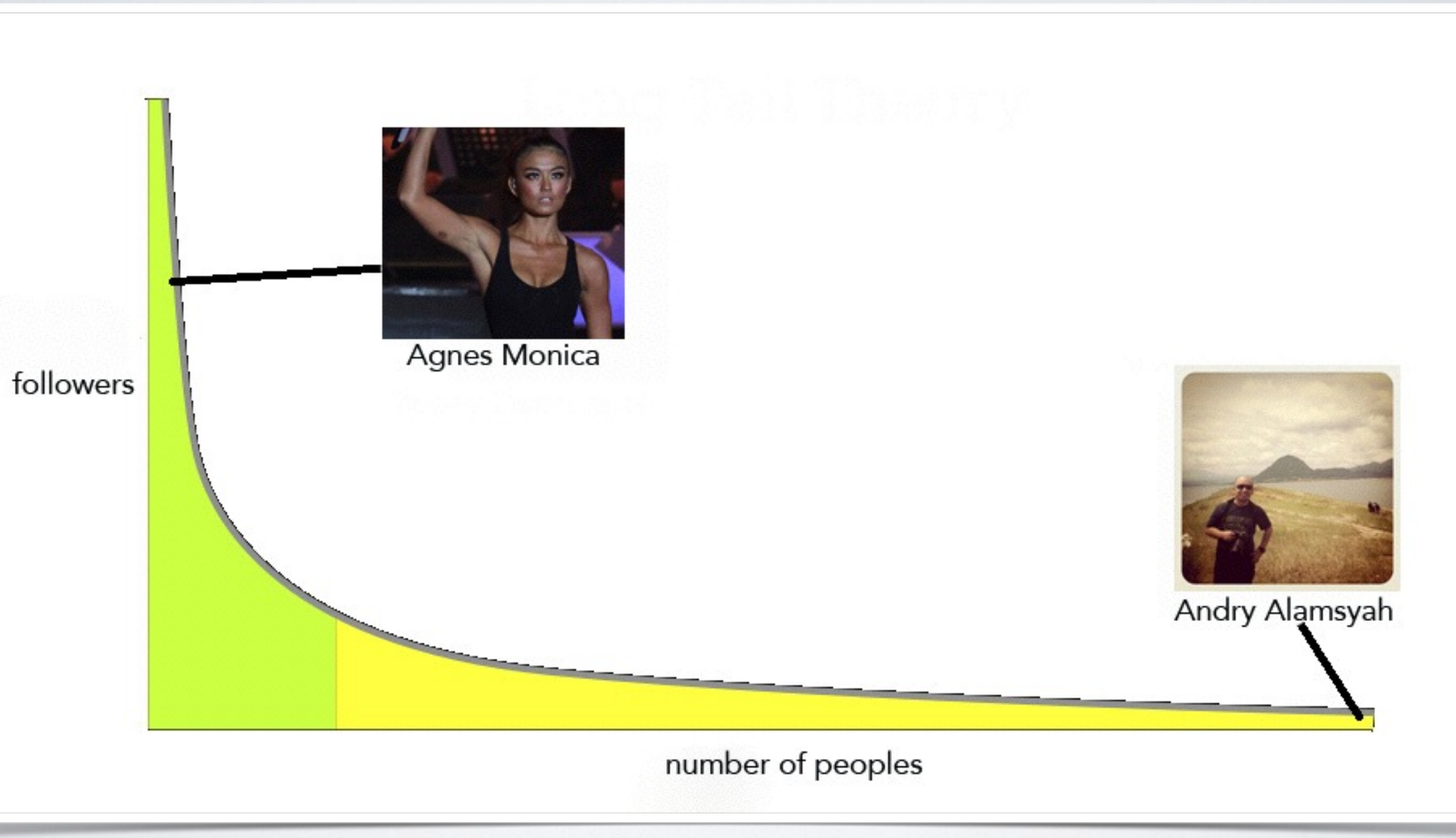


SOCIAL NETWORK CHARACTERISTICS



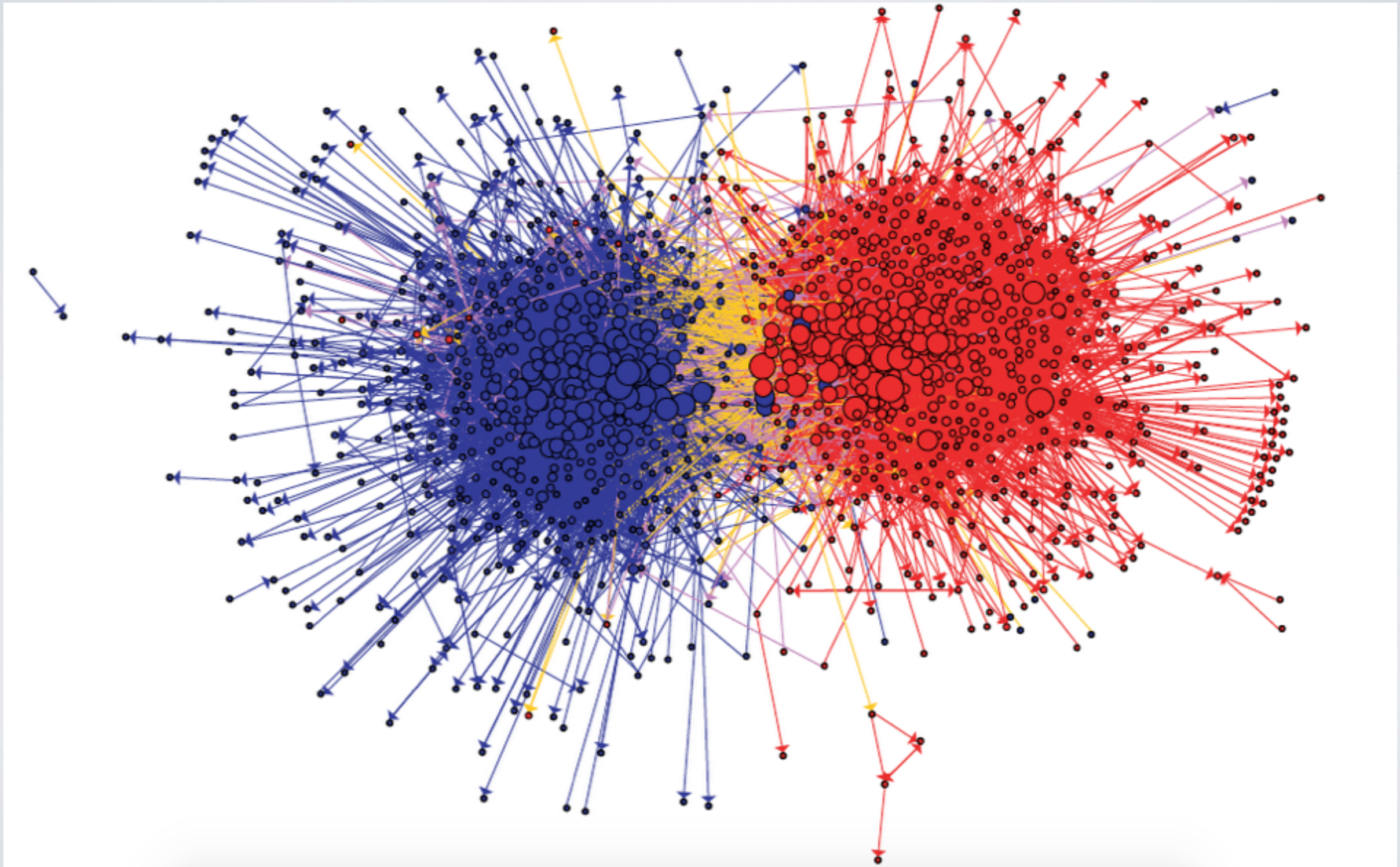


SOCIAL NETWORK CHARACTERISTICS



power law distributions

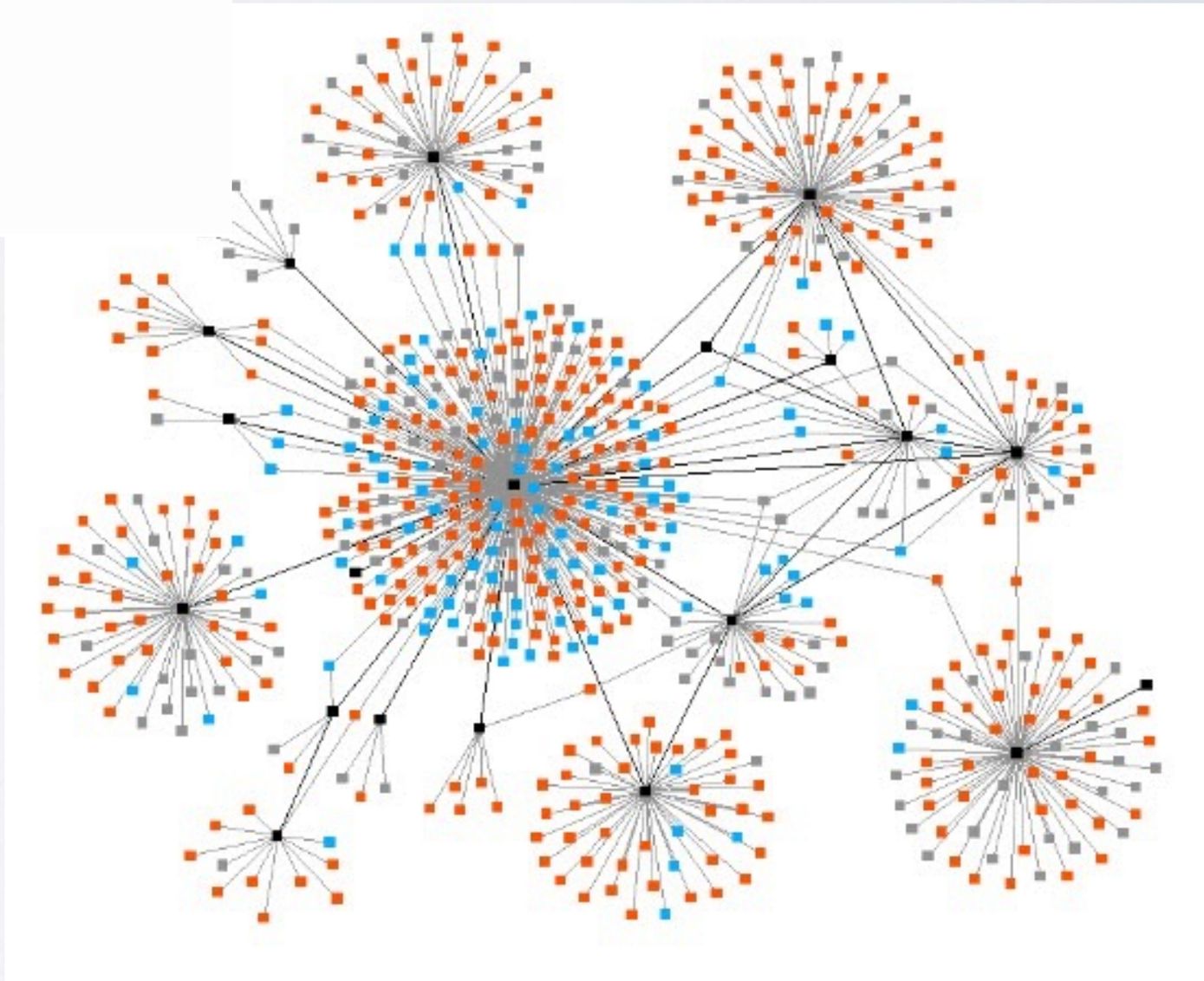
SOCIAL NETWORK CHARACTERISTICS



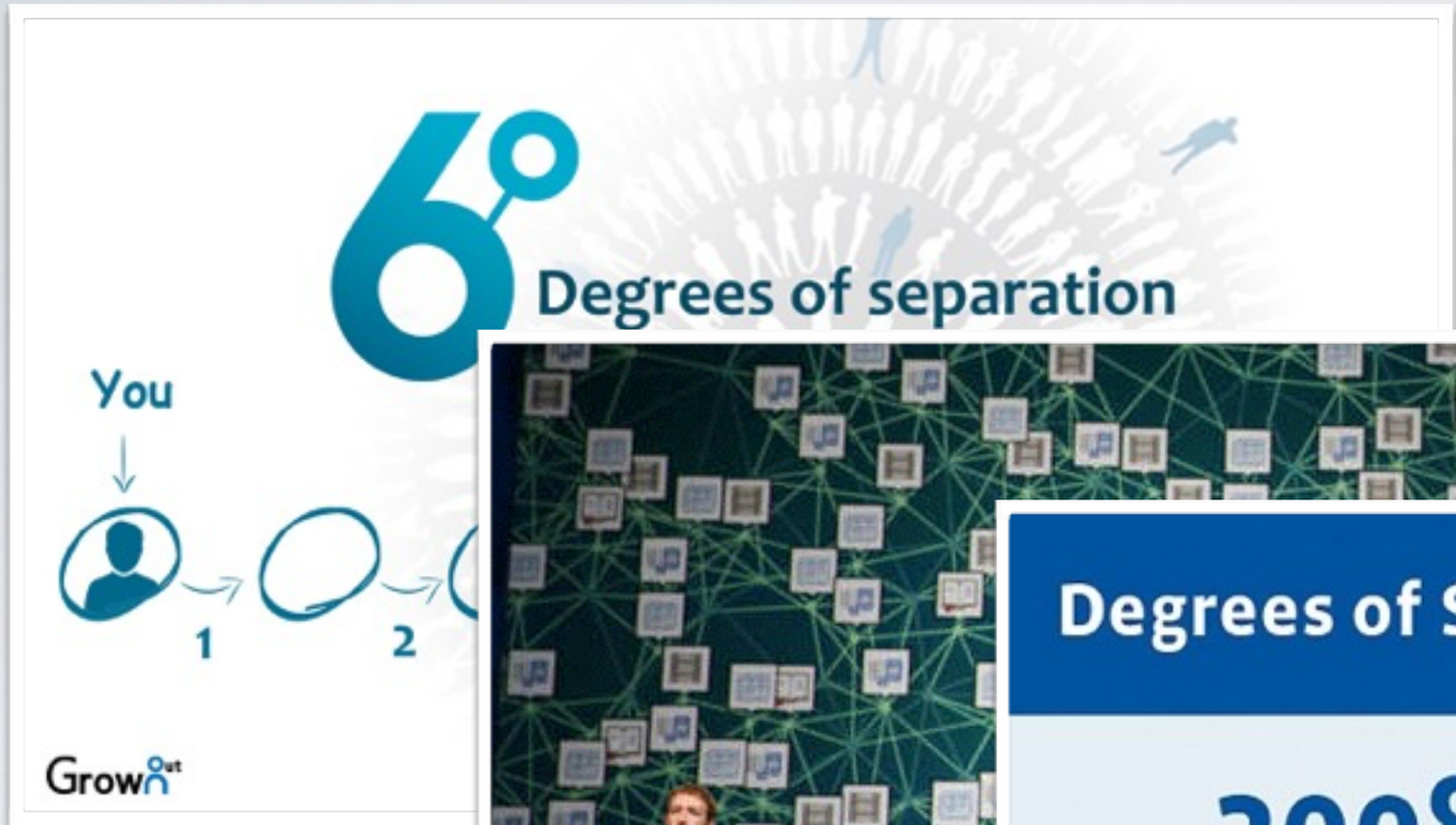
SOCIAL NETWORK CHARACTERISTICS

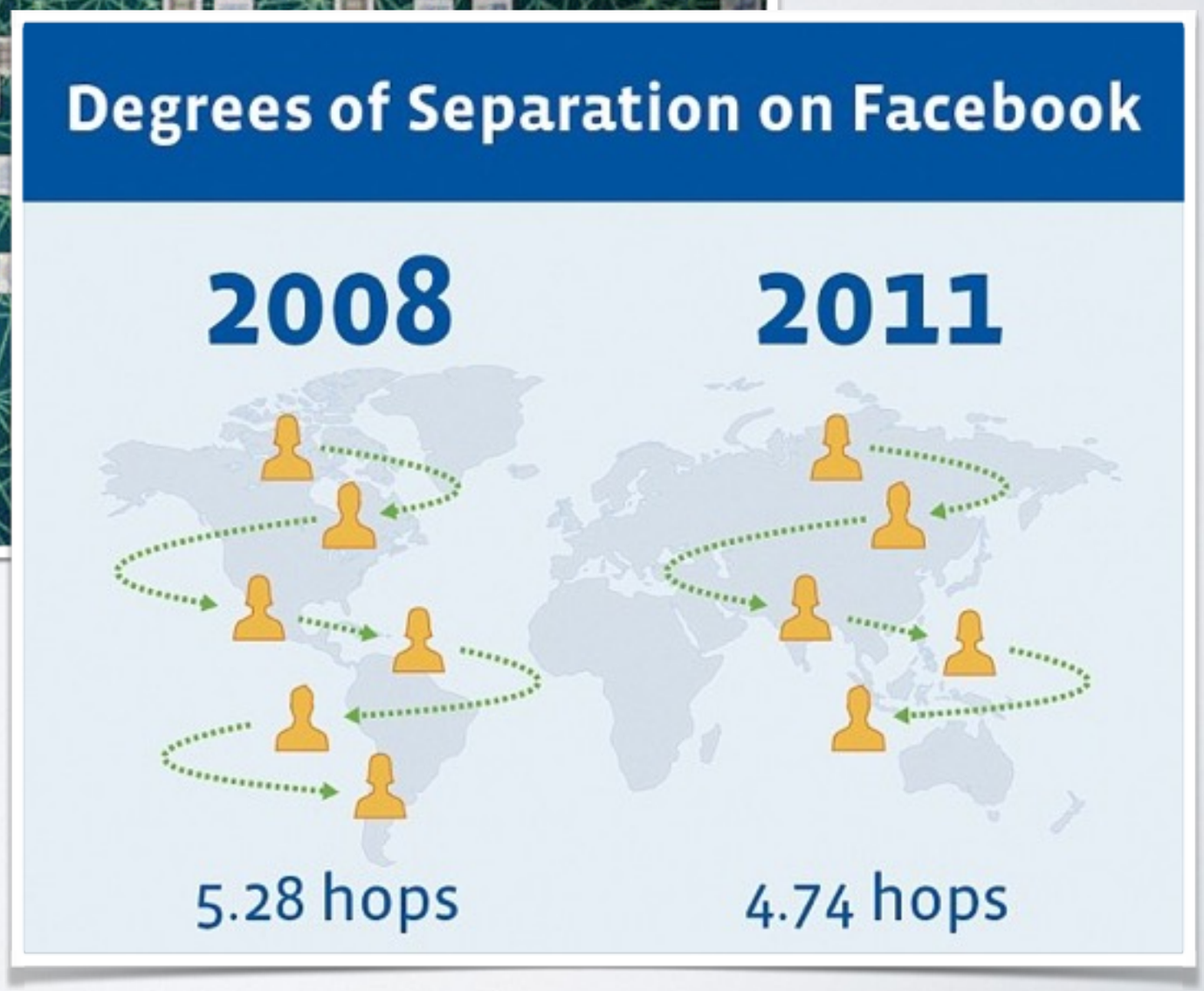
growth & preferential attachment

small world



SOCIAL NETWORK CHARACTERISTICS



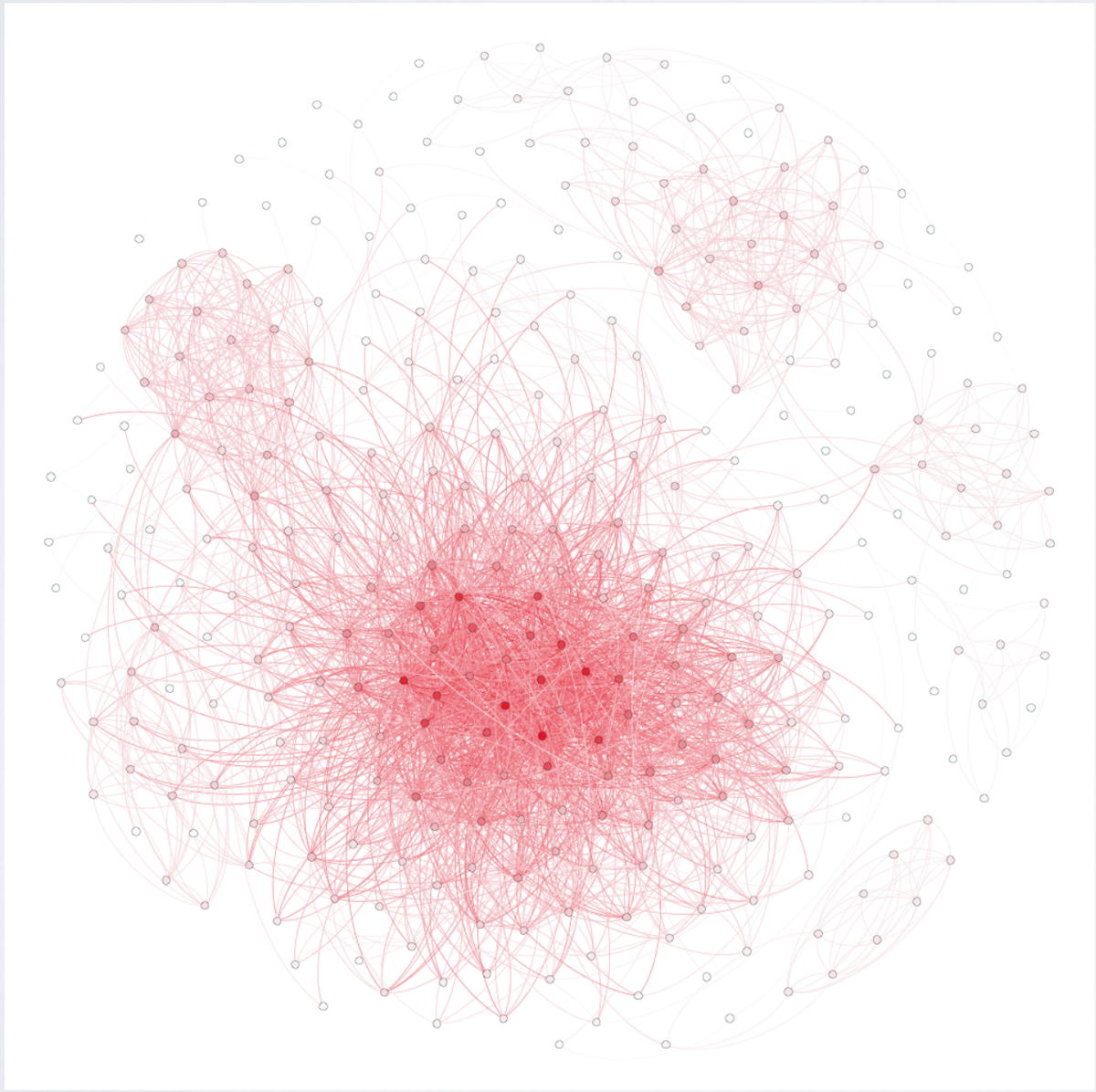


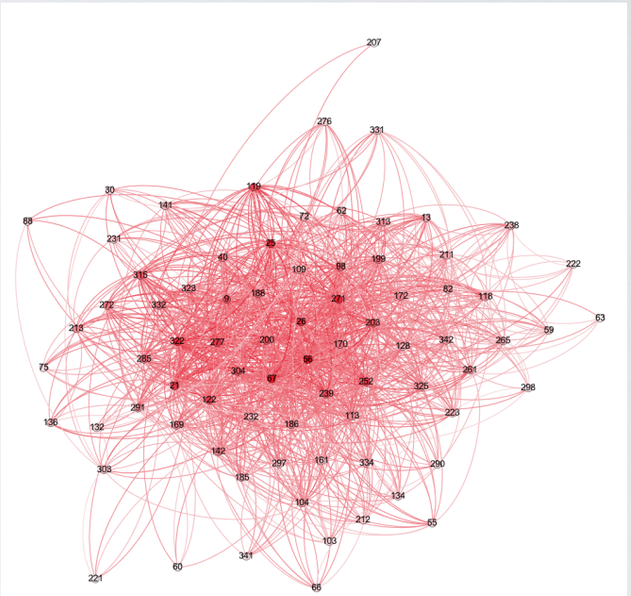
SOCIAL NETWORK INTERPRETATIONS

Case	*Questions*	*SNA Tools*
Leader Selection	Who is the central in the trust and respect network?	Degree Centrality
Ranks	How do we rank our top performer individuals in the organization?	Eigenvector Centrality, Pageranks
Task Force Selection	How do we put together a team that maximally connected through out the	Closeness Centrality
Mergers and Acquisition	How to merge separate cultures / networks?	Homophilly, Reciprocity, Mutuality, Transitivity
Competitive Advantages	What is the missing links between supply and demand?	Structural Holes
Advertising Attachment	How strong the impact of our advertisement effort?	Tie Strength
Market Segmentation	How segmented our market is?	Clustering Coefficient, Clique, Cohesive, Modularity
Information Dissemination	How is the information / knowledge spreading?	Random Walks, Hits Algorithm
Strength out the organization	How to increase redundancy and interconnectedness?	Bridge
Dynamics of Organization	How dynamics our organization is?	Temporal Networks

PUBLICATIONS (1)

Financial Fraud Detection using Social Network Analysis (eii-Forum, 2013)





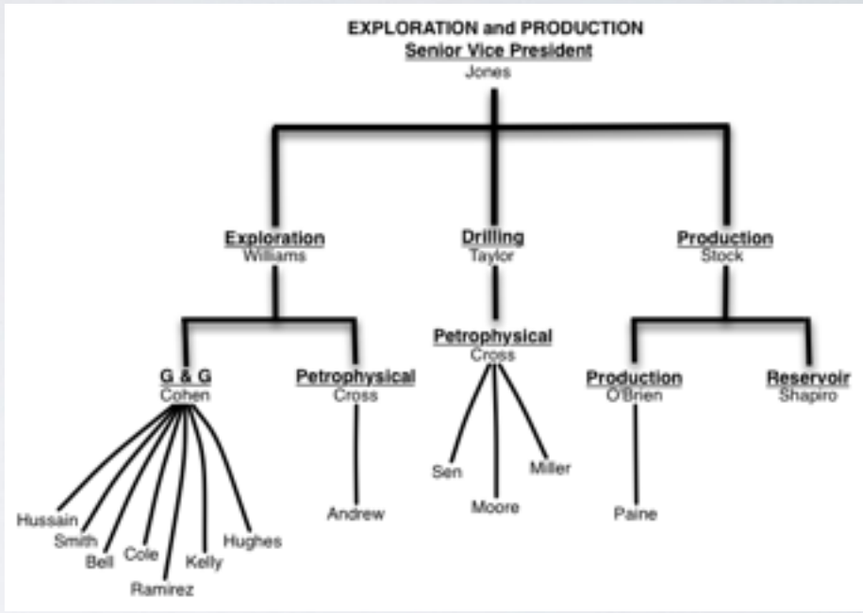
Tabel 3. komunitas, persentase dan node dengan degree terbesar pada komunitas.

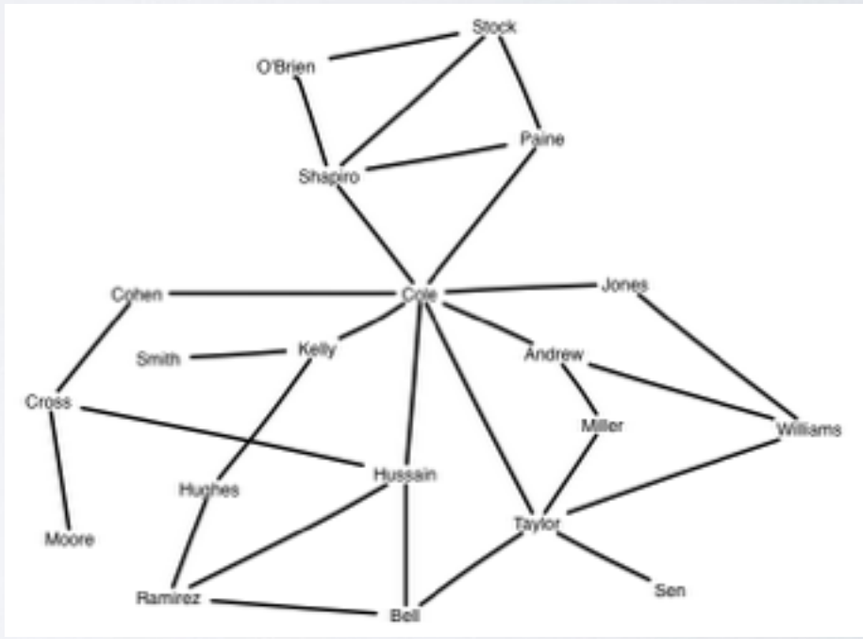
Komunitas	*Persentase (%)*	*Node dengan Degree Terbesar (a)*	*Degree Node (a)*
2	31,23	56	77
7	13,51	312	25
3	12,61	175	16
1	12,31	271	72
0	10,21	53	30
4	8,11	119	61
5	3,9	320	20
8	3	275	9
10	2,4	89	7
9	0,9	179	2
12	0,6	233	1
11	0,6	282	1
6	0,6	33	1

PUBLICATIONS (2)

The Role of Social Network Analysis in Knowledge Management (Jurnal Manajemen Indonesia, 2013)

Case	*Questions*	*SNA Tools*
Leader Selection	*Who is the central in the trust and respect network?*	*Degree Centrality*
Ranks	*How do we rank our top performer individuals in the organization?*	*Eigenvector Centrality, Pageranks*
Task Force Selection	*How do we put together a team that maximally connected through out the organization?*	*Closeness Centrality*
Mergers and Acquisition	*How to merge separate cultures / networks?*	*Homophilly, Reciprocity, Mutuality, Transitivity*
Competitive Advantages	*What is the missing links between supply and demand?*	*Structural Holes*
Advertising Attachment	*How strong the impact of our advertisement effort?*	*Tie Strength*
Market Segmentation	*How segmented our market is?*	*Clustering Coefficient, Clique, Cohesive*
Information Dissemination	*How is the information / knowledge spreading?*	*Random Walks, Hits Algorithm*
Strength out the organization	*How to increase redundancy and interconnectedness?*	*Bridge*
Dynamics of Organization	*How dynamics our organization is?*	*Temporal Networks*

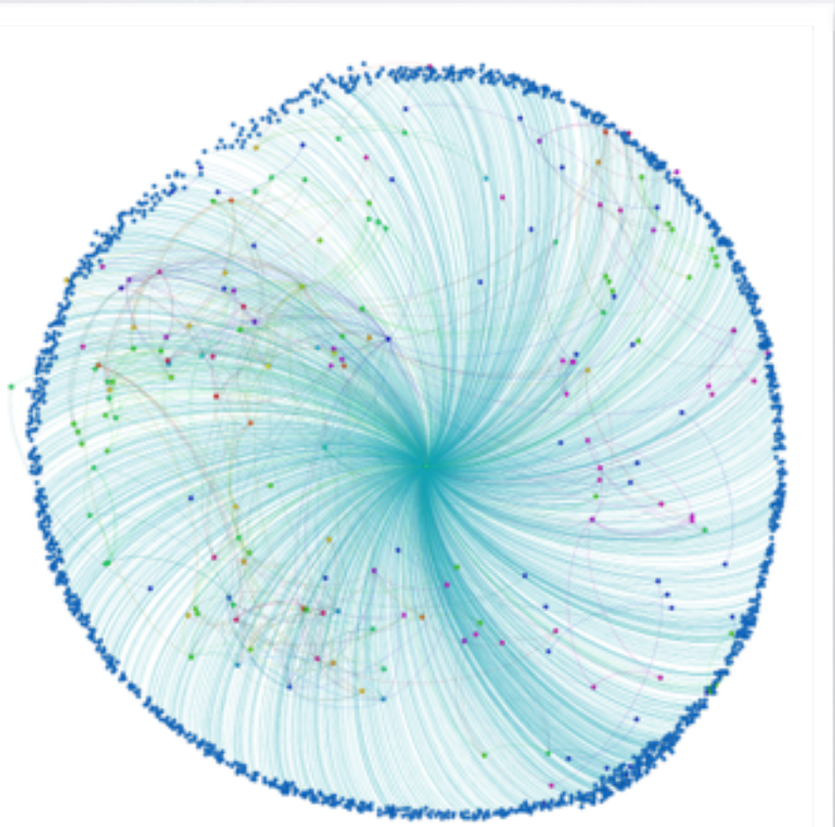


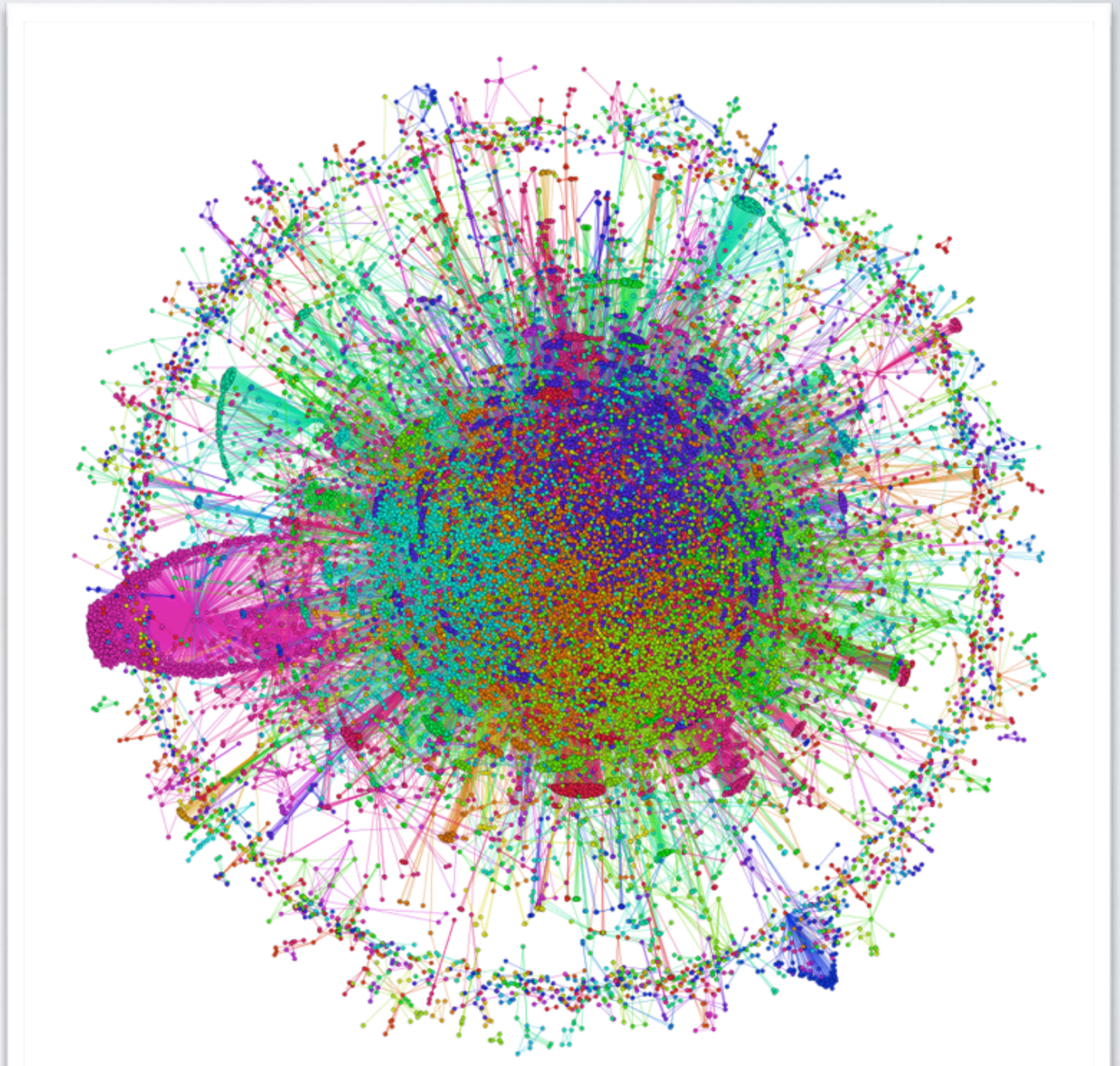


Vizualisation of hierarchical structure organization and knowledge flow of informal organization (Alamsyah, 2013)

PUBLICATIONS (3)

Effective Knowledge Management using Big Data and Social Network Analysis (ISCLO, 2013)





PUBLICATIONS (4)

Social Network Modelling Approach for Brand Awareness (ICOICT, 2014)

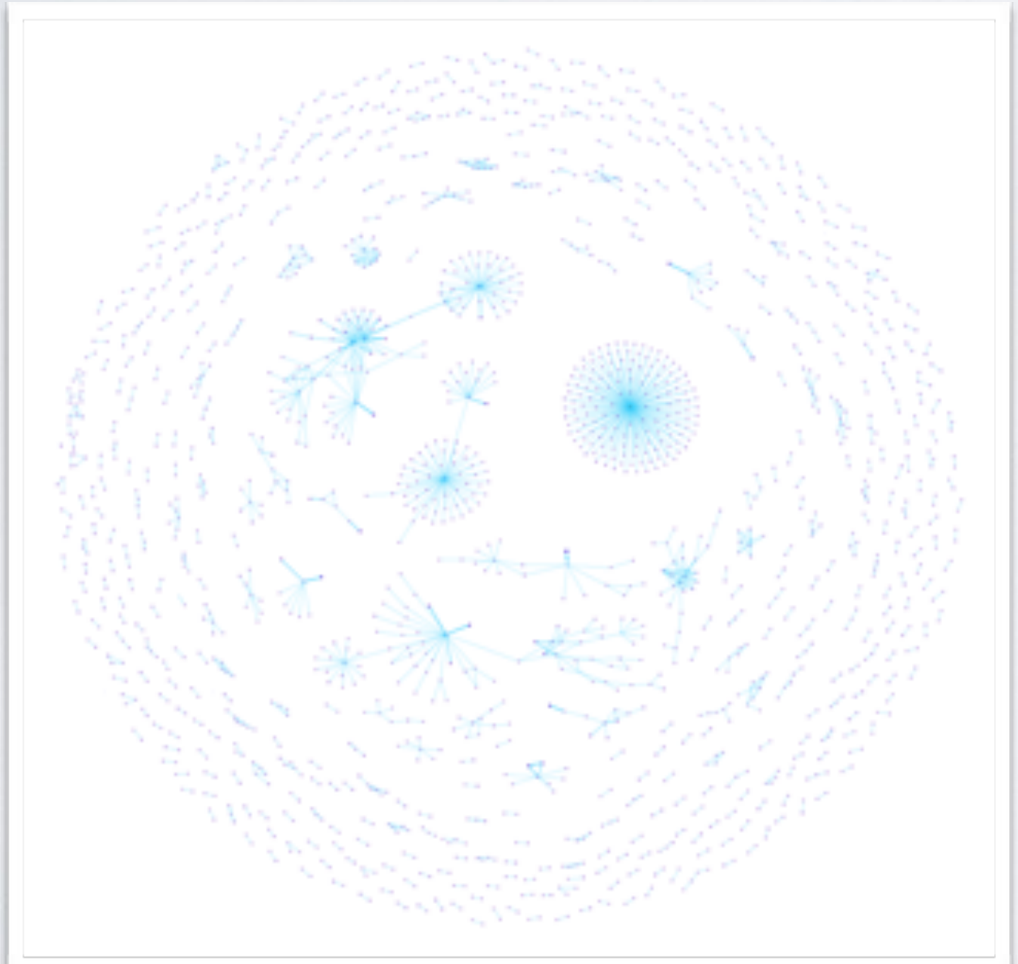


Figure 5. The graph model of twitter keyword "Telkom"

TABLE III. CENTRALITY RANK

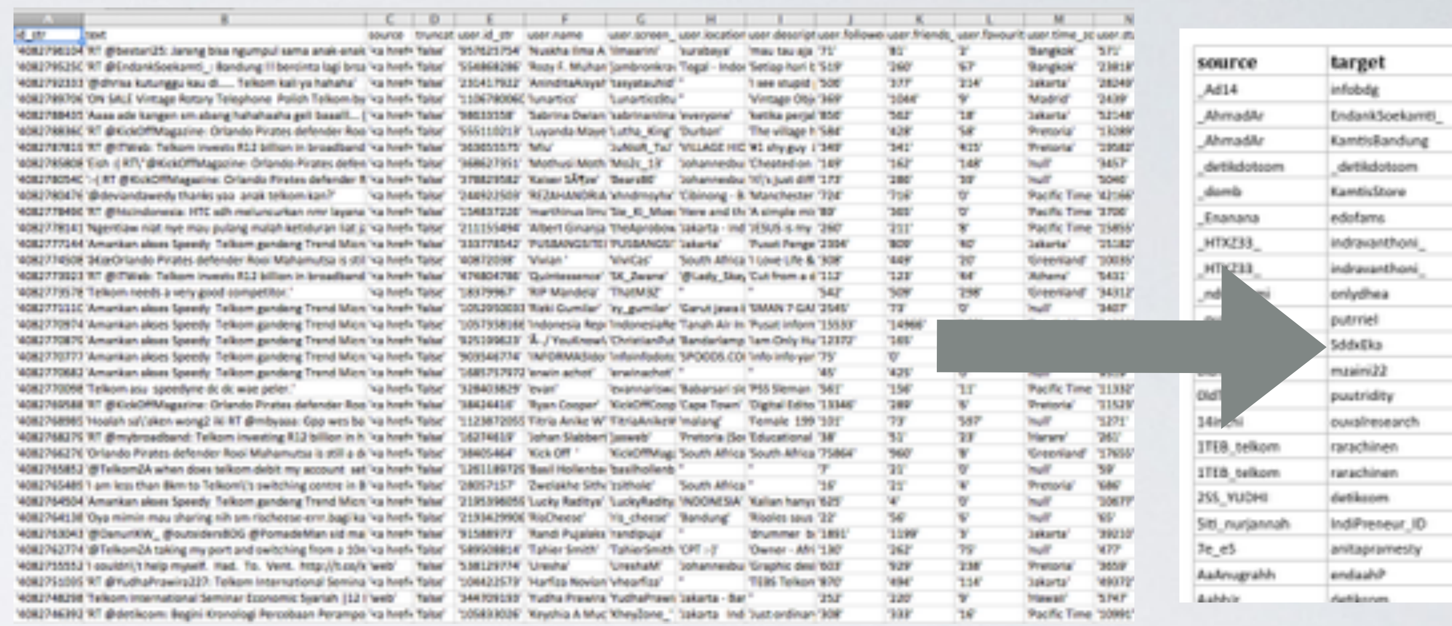
No	Account	DC (value/rank)	BC (value/rank)	Community
1	@ouvalresearch	152 / 1	0.006 / 1	7
2	@detikcom	55 / 2	0.008 / 8	9
3	@infobdg	51 / 3	0.003 / 2	0
4	@EndankSoekamti	43 / 4	0.002 / 3	1
5	@SID_Merch	42 / 5	0.002 / 4	1

TABLE II. COMMUNITIES RANK

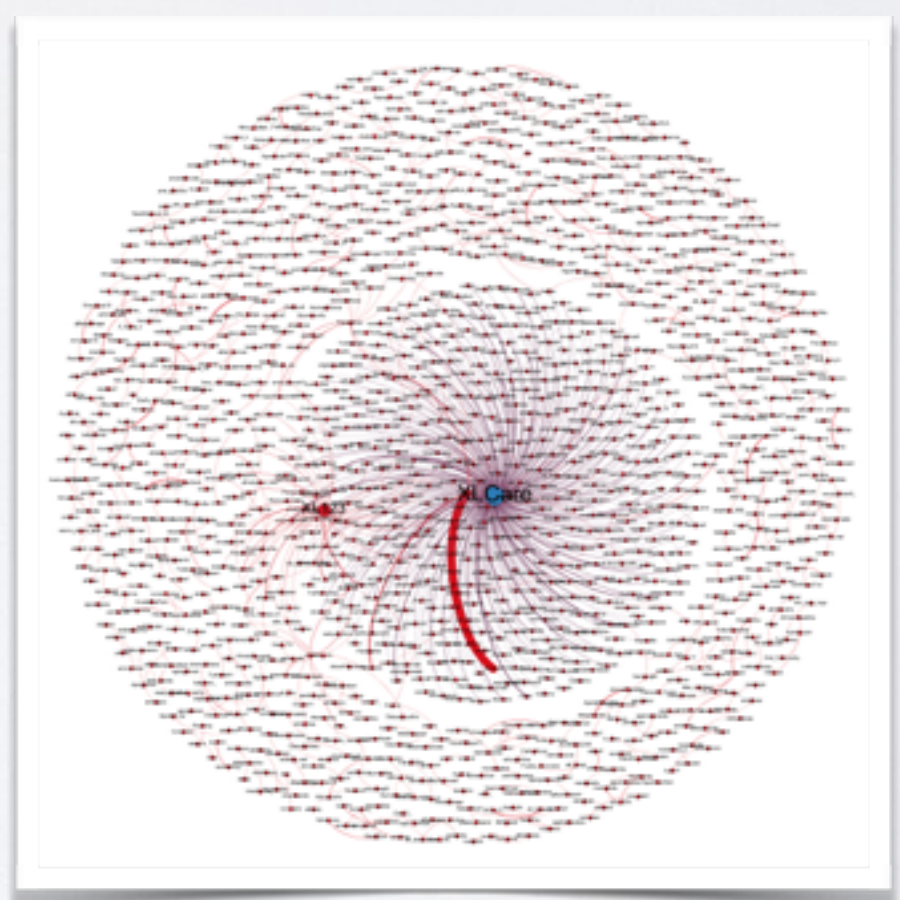
No	Community Number	Number of Node	Size (Percentage)
1	7	153	7.55
2	1	133	6.57
3	0	79	3.90
4	9	78	3.85
5	41	55	2.71
6	62	37	1.82
7	72	34	1.67
8	35	28	1.38

PUBLICATIONS (5)

Topic : *Social Customer Relationship Management based on Social Conversational Data* (ISCLO, 2014)

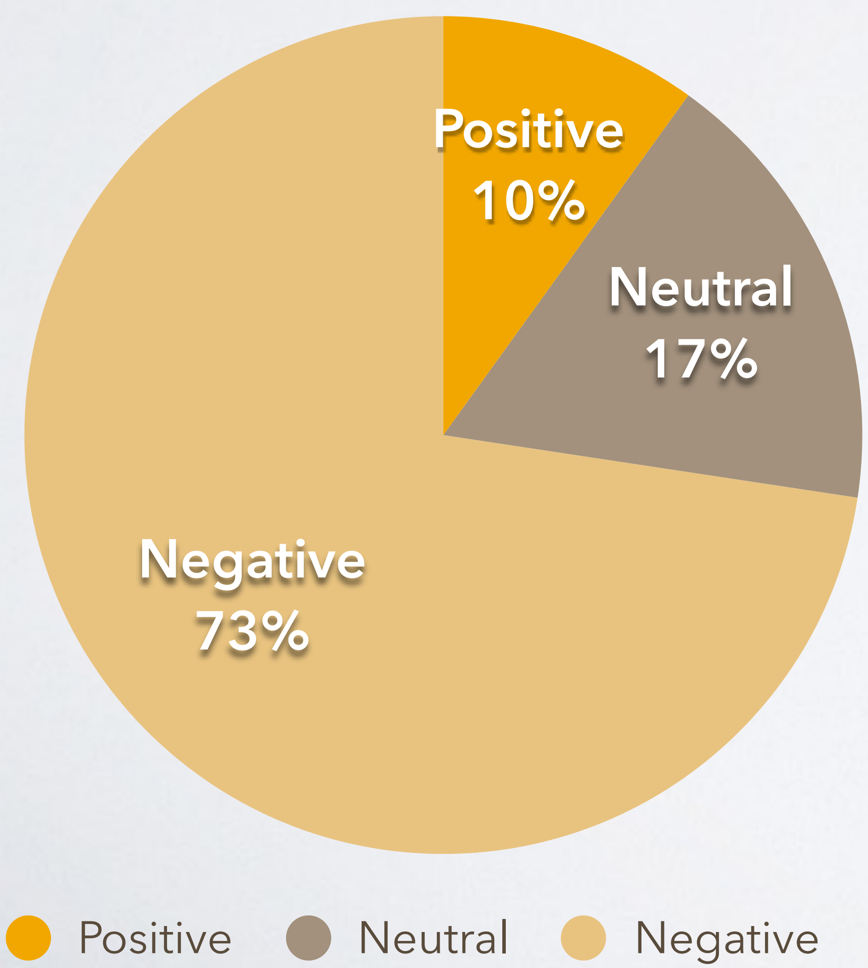


transforming conversation data into social network



conversation network around keyword & hashtag of "XL", "XLCare", "XL123"

SENTIMENT ANALYSIS



PUBLICATIONS (6)

Network Market Analysis using Large Scale Social Network Conversation in Indonesia's Fast Food Industry (ICOICT, 2015)

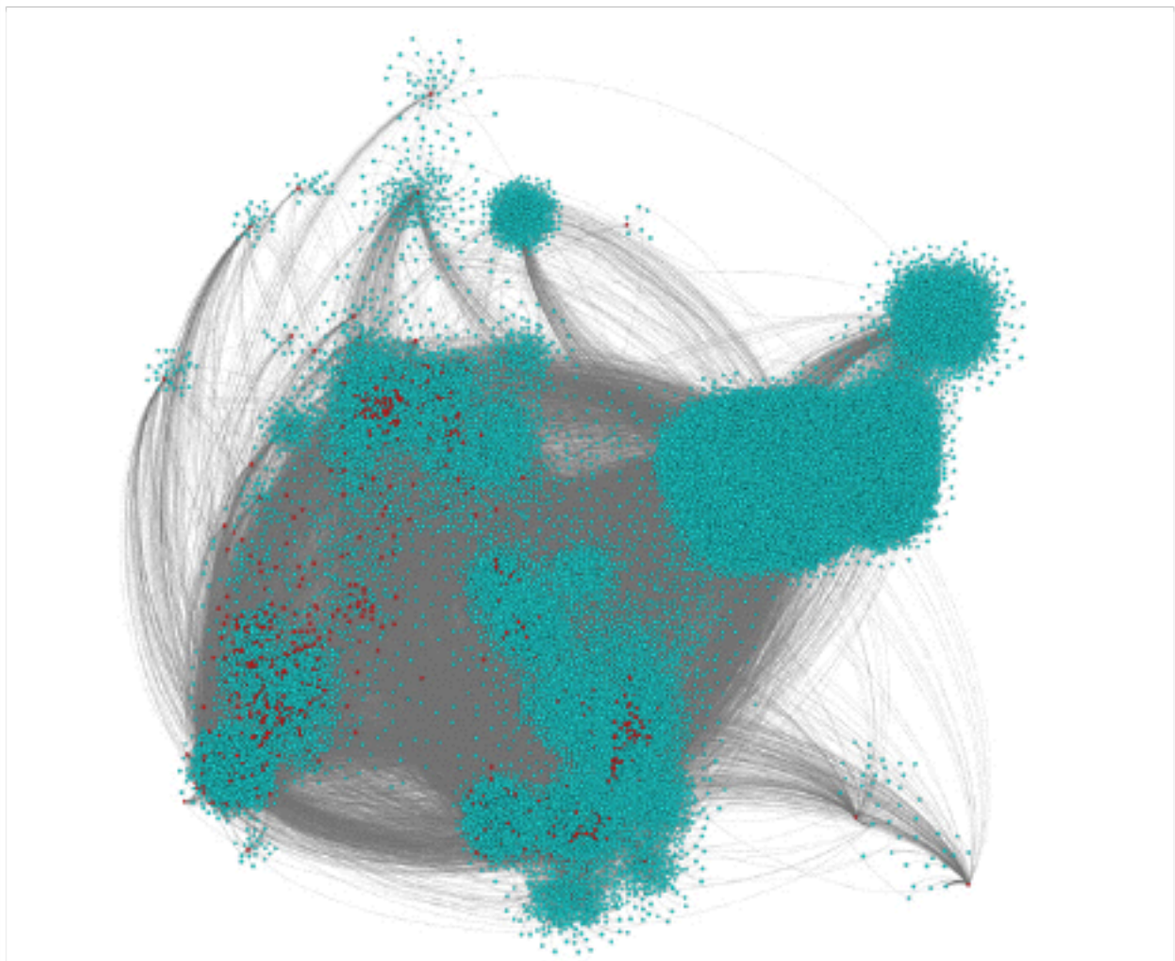


Fig. 2. Large scale MD network, green colour are user posts, while other colour are page posts.

TABLE I. SNA METRICS COMPARISON

No	Metric	MD Network	BK Network
1	Size (Number of Likes)	55,423,739	107,102
2	Size (Bipartie Network)	438,568 nodes 909,748 edges	220,289 nodes 226,337 edges
3	Average Degree	2.418	0.116
4	Density	0.00001097725	0.000003890513
5	Average Path Length	3.42506	3.245614
6	Diamater	8	8
7	Modularity	0.670 (13 communities)	0.527 (10 Communities)

TABLE III. FURTHER INFORMATION ABOUT THE NETWORKS

No	Description	MD Network	BK Network
1	POST TYPE	99.7% USER COMMENTS, THE REST ARE VIDEO, STATUS, PHOTO, LINK	99.55% USER COMMENTS, THE REST ARE VIDEO, STATUS, PHOTO, LINK
2	5 HIGHEST ENGAGEMENT VALUE	137609, 132372, 113299, 99668, 97973	129423, 25625, 25053, 21295, 17046
3	POST TYPE OF 5 HIGHEST ENGAGEMENT VALUE	PHOTO, STATUS, STATUS, LINK, PHOTO	ALL PHOTO
4	USER SEX	55.27% MALE, 41.50% FEMALE	55.68% MALE, 43.83% FEMALE

PUBLICATIONS (7)

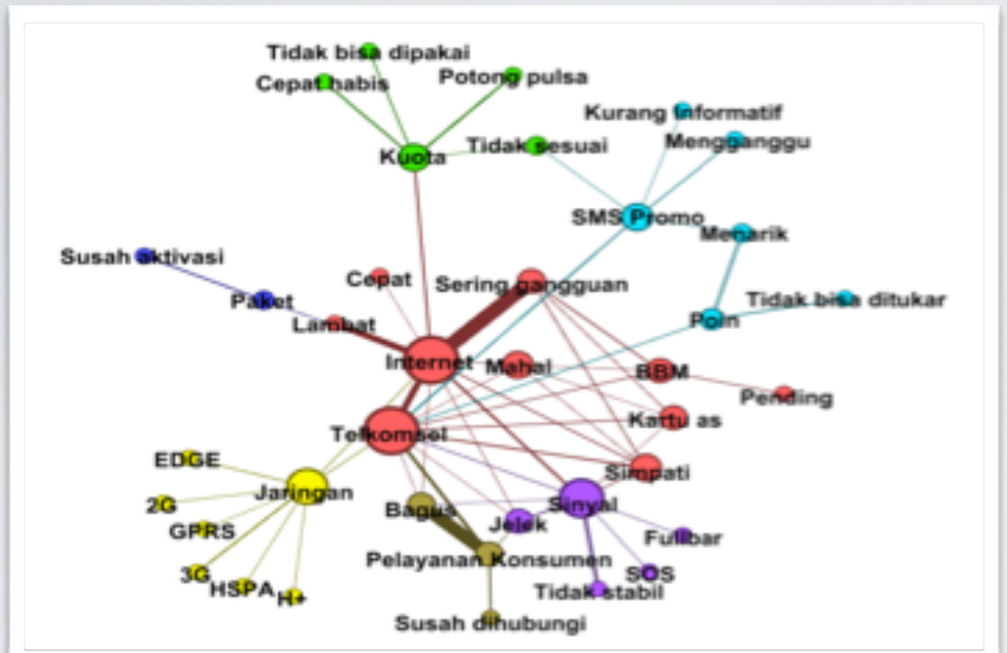
Indonesian Music Fans Group Identification using Social Network Analysis in Kaskus Forum (ICOICT, 2015)

Fig. 8. Network visualization shows only the groups which has 3 members or more

TABLE III. SIZE DISTRIBUTION PERCENTAGE

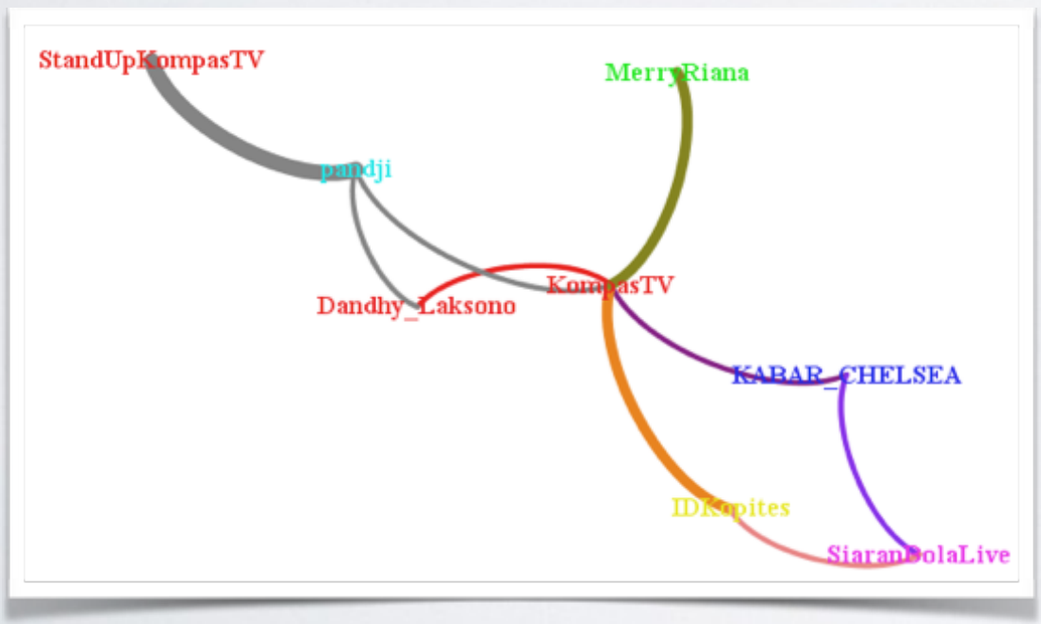
Size	The Number of Groups	Percentage
123	1	9,34%
115	1	8,73%
78	1	5,92%
76	1	5,77%
56	1	4,25%
50	1	3,8%
49	1	3,72%
45	1	3,42%
40	1	3,04%
15	1	1,14%
3	1	0,23%
2	8	0,15% x 8
Total	**19**	**50.56%**

UNPUBLISHED WORK

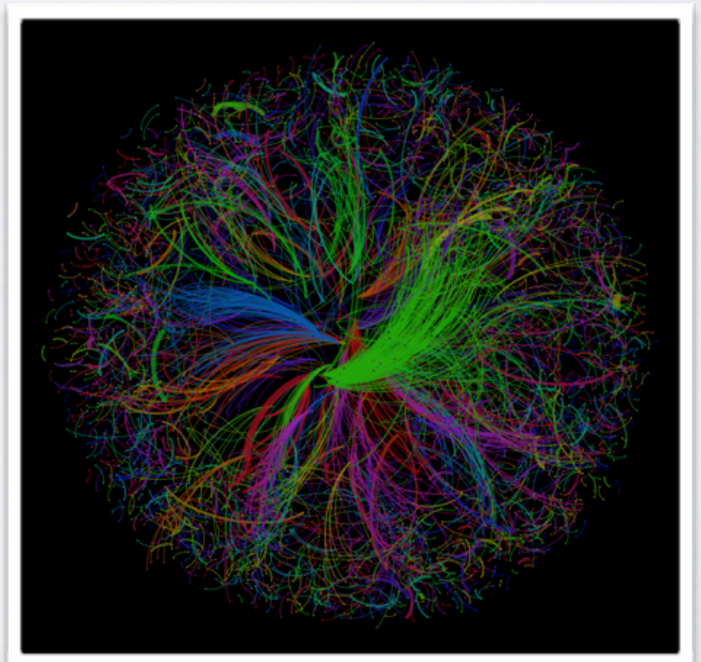


Social Network Analysis, Text Mining, Sentiment Analysis for CRM (Telkomsel)

Telkom University Academics Network (academia.edu)



Influencer Analysis for Alternative Marketing strategy of KompasTV



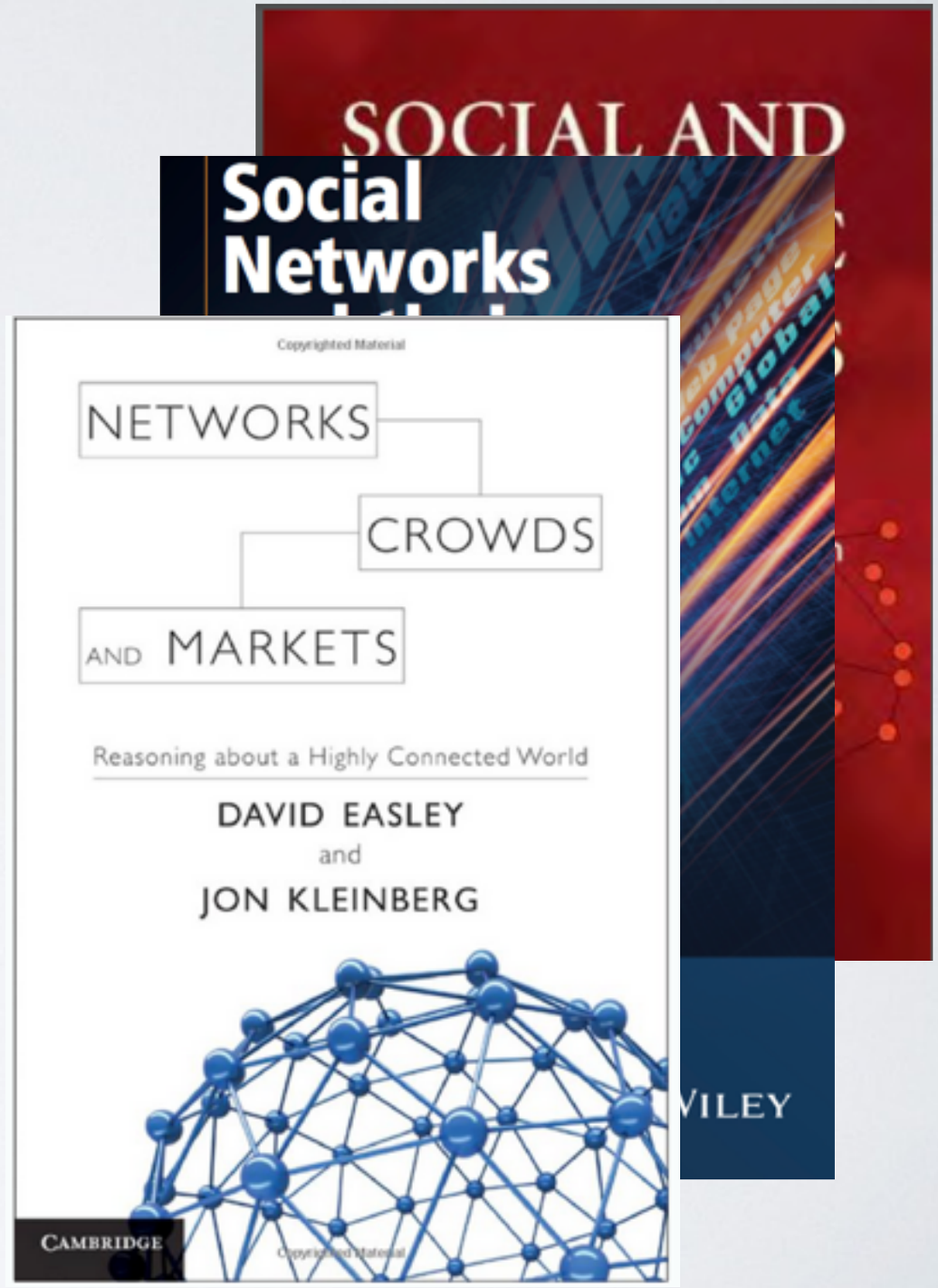
Word of Mouth Simulation Process on Hashtag #Android #BBM

ONGOING RESEARCH

- *Social Engagement Analysis* in online conversation of higher education community; use *Facebook* data
- *Top of Mind (Brand Awareness)* methods comparison using legacy approach and social media crawling data; verification research
- *Sentiment Analysis* using appraisal theory; mobile phone product comparison
- *New Metric Constructions* for engagement and dissemination information (example: marketing purpose)
- *Smart City Metric* based on social media
- Classification models for online SME selling and production effort
- Data mining modelling for online advertising and online tourism
- Indonesian investor and start up network model. Based on small world phenomenon
- Text network analysis comparison for telco industry
- Online conversation comparison of three biggest bank in Indonesia
- etc …

OPPORTUNITY

- dynamics network (size, movement, temporal, evolving network)
- network market predictions
- social network aggregate behaviour (game theory on traffic, market, auctions)
- bargaining and powers in networks
- social and economic network (strategic formation, diffusion, learning from networks, etc)
- social network and communication pattern in the context of smart city
- metric construction
- graph algorithm
- etc ..



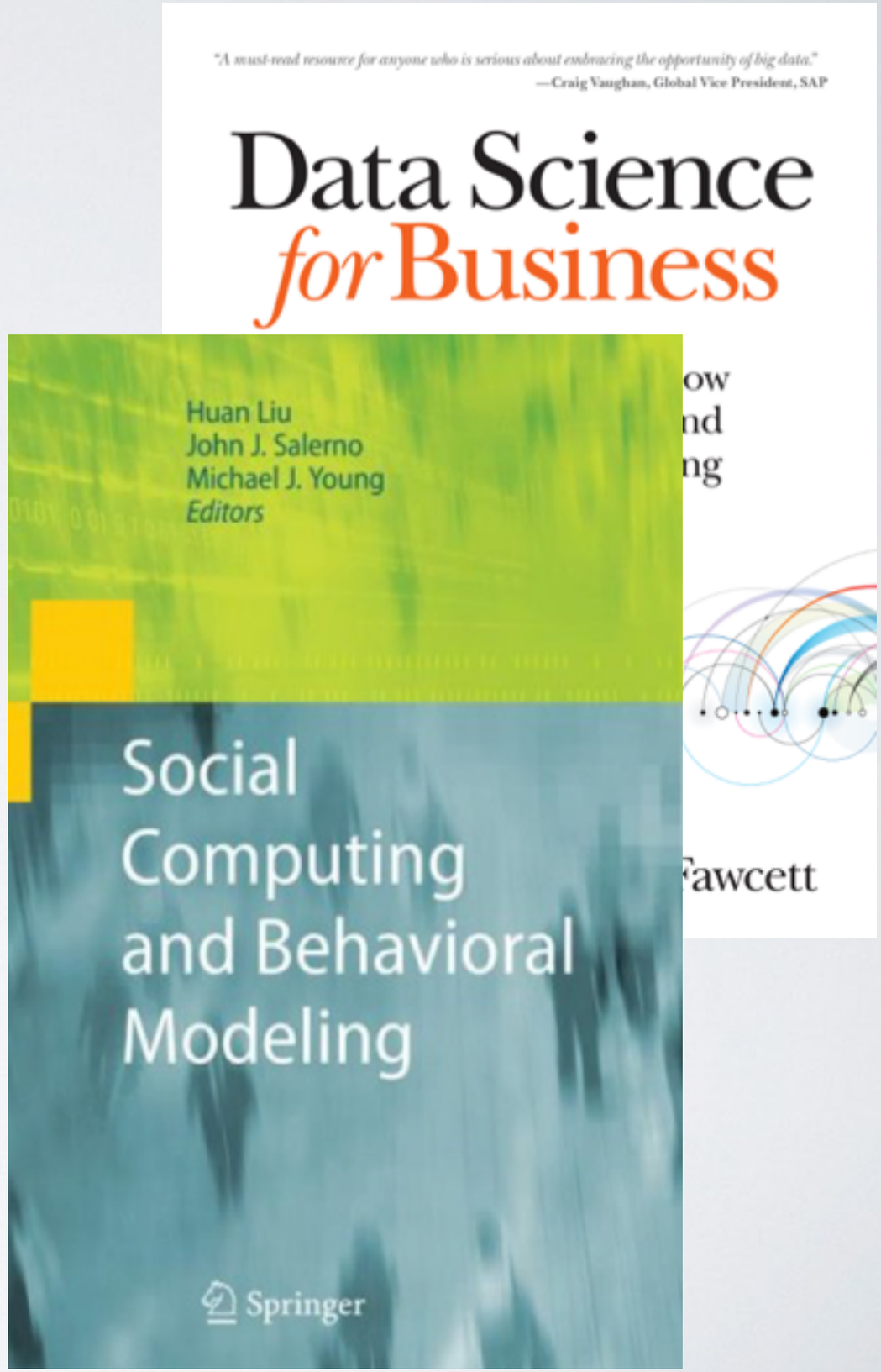
CONCLUSIONS

- The availability of large volume dataset should encourage business and social science (as well as other sciences) to apply *Big Data* (SNA/DM/KDD/DS) tools

- Business and Social Science (as well as other sciences) need to use more *Big Data* tools, because of the capability to model, predict, simulate and optimise phenomenon

- There are many vacancy in Social Network (Graph Representation) research, such as large scale graph, graph reduction, graph algorithm

- *Big Data* approach are more and more utilised as standard of decision making in modern business



without Big Data, you are blind and deaf in the middle of a freeway - Geoffrey Moore

TERIMA KASIH