BITCOIN NETWORK AND MACHINE LEARNING

Cazabet Rémy

WHO AM I

- Rémy Cazabet
- Associate Professor (Maître de conférences)
  - Université Lyon 1
  - LIRIS, DM2L Team (Data Mining & Machine Learning)
- Computer Scientist => **Network Scientist**
- Member of IXXI, Lyon's institute of **Complex Systems**

SHORT NOTICE

- **Launch Gephi download**
  - https://gephi.org/users/download/
- Course content, resources…
  - http://cazabetremy.fr/Teaching/BitcoinNetwork.html
  -

COMPLEX SYSTEMS

- **Complex systems**:
  - Systems composed of multiple **parts** in **interactions**
  - The **macro level** behavior of the system depends on the **micro level**, and reciprocally (≠ micro, macro in economics…)
- Interdisciplinary field, thought to solve the problems of the **reductionist** approach.
  - Reductionism: to understand a system, we need to understand its parts
  - Complex system: we need to understand how parts are interacting.

EXAMPLE OF CS

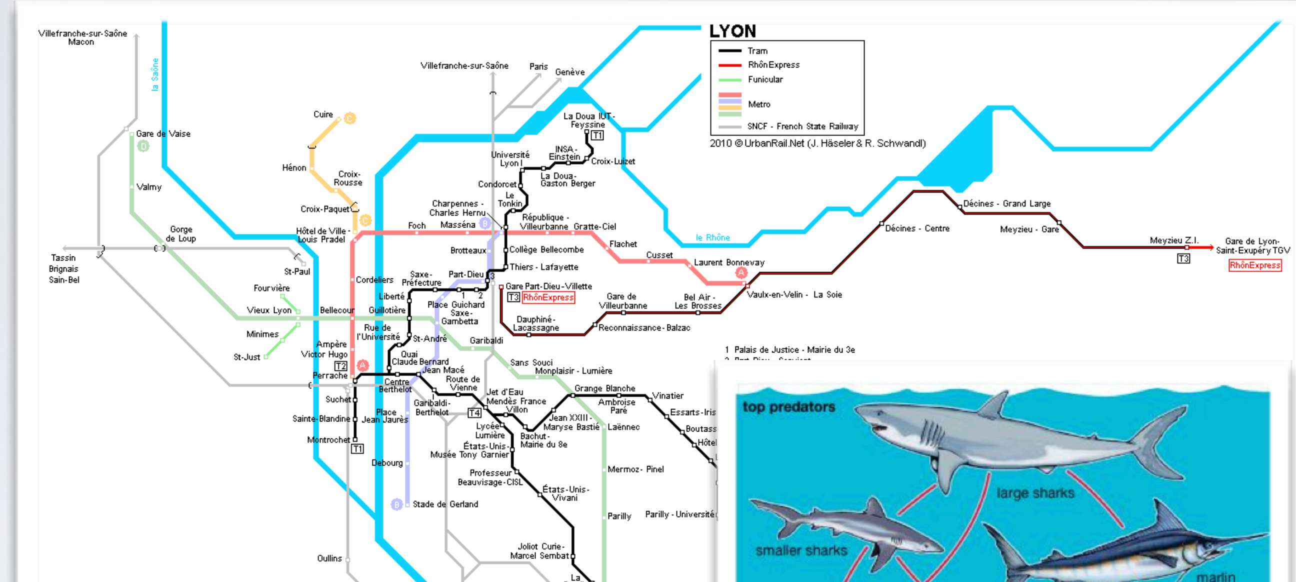
- Typical complex systems:
  - Organisations, cities, human body, brain, ecosystems, etc.
- Brain:
  - Micro level: neurons, synapses, receptors (light, sound, …), chemicals, …
  - Macro level: Signals, synchronization, … => Intelligence, decisions…
- Social networking platforms:
  - Micro level: individuals, companies, bots, hackers, posts, communications…
  - Macro level: information diffusion, patterns of activity, echo chamber/filter bubble, fake news, rich get richer phenomenon, etc.

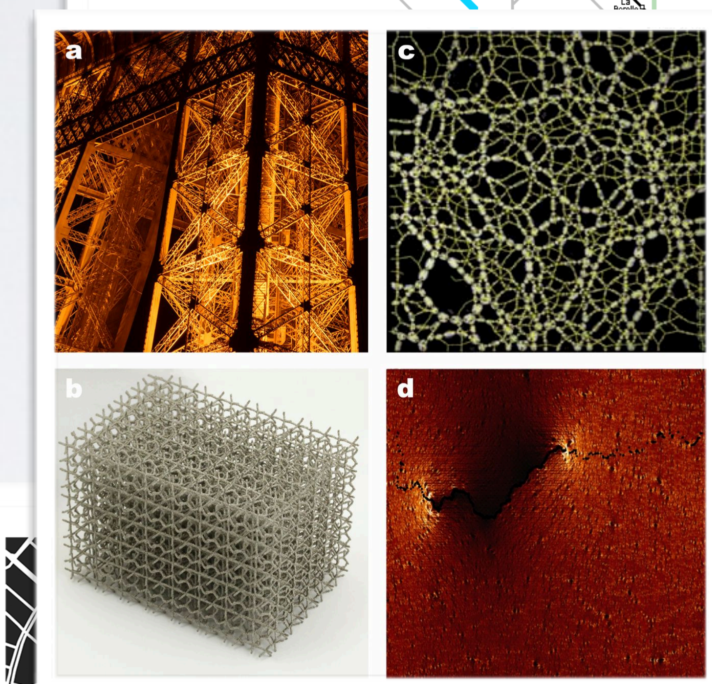
EXAMPLE OF CS

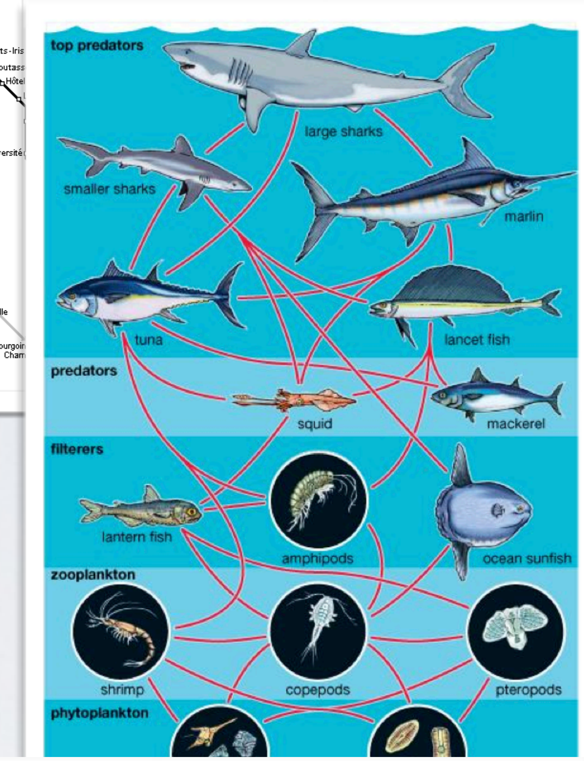
- Economy ? (Financial) Markets ?

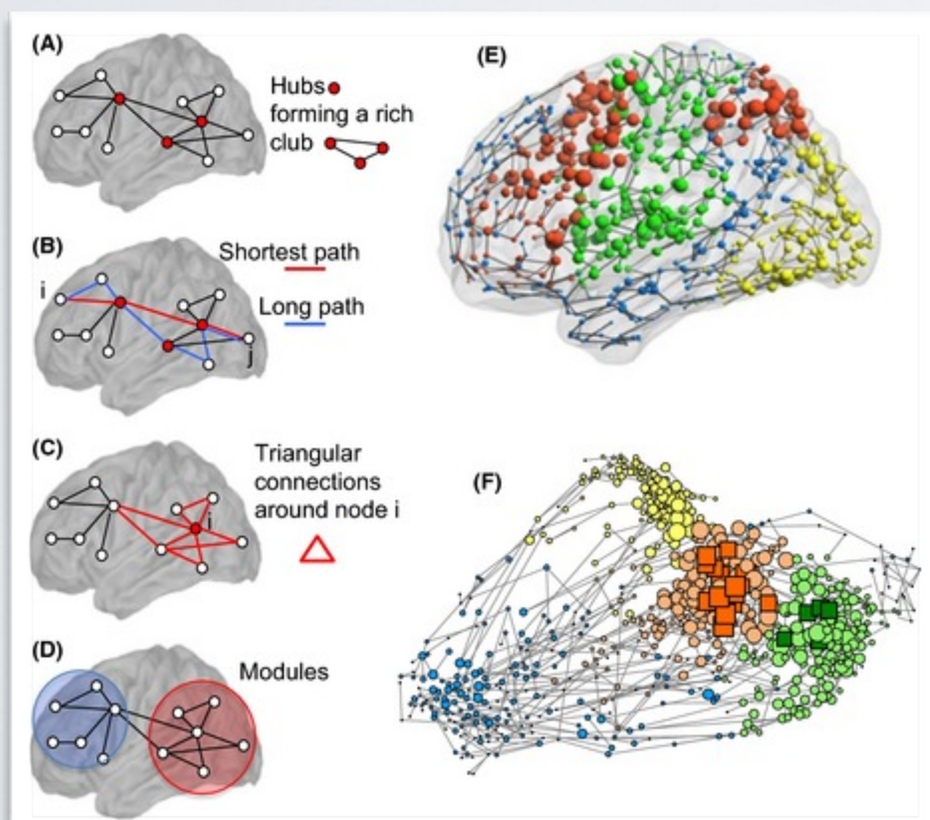
NETWORK SCIENCE

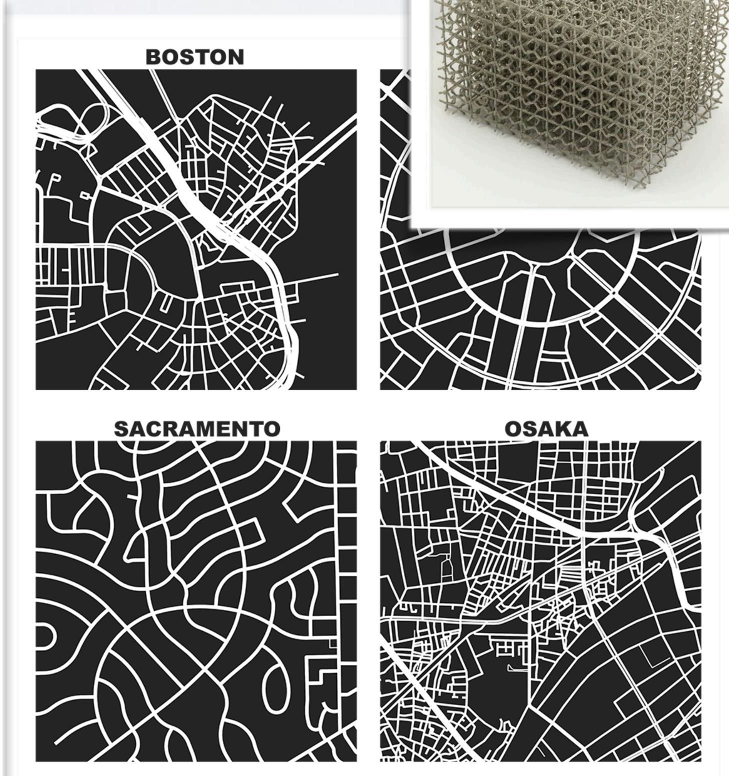
- Study interactions between entities at the micro level => represent interactions as a **network**
- Analyse this network based on tools from **network science**
- Vocabulary: network science ≈ Complex/Social network analysis ≈ Graph mining

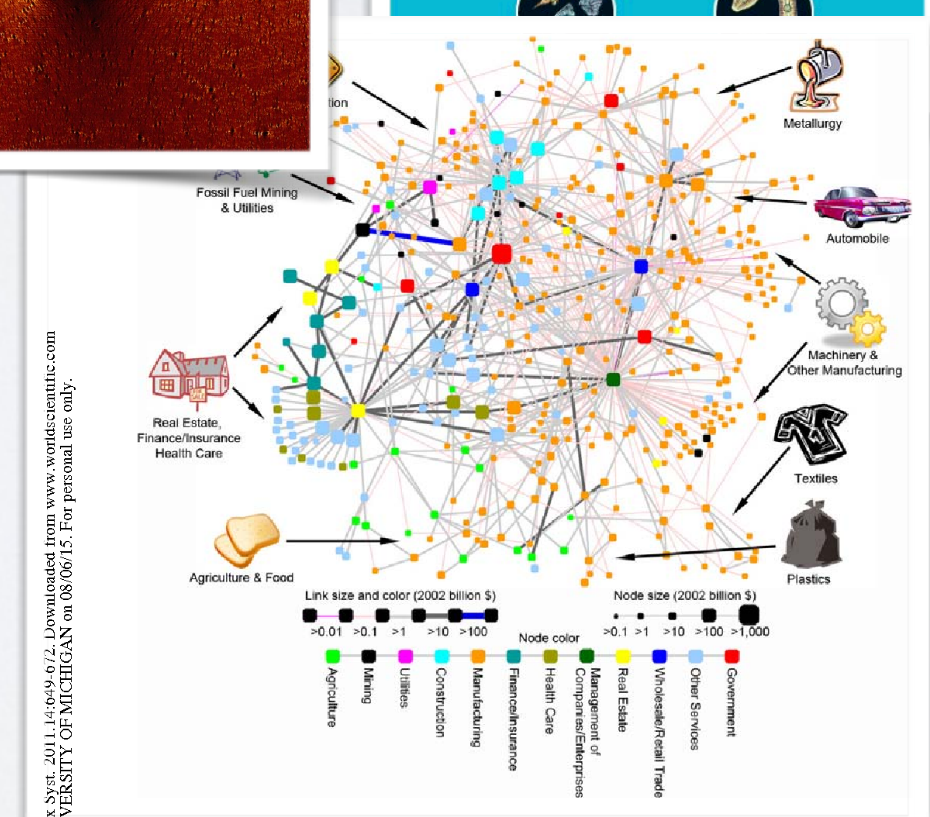

LYON
Tram
RhônExpress
Funicular
Metro
SNCF - French State Railway
2010 © UrbanRail.Net (J. Häseler & R. Schwandl)


a
c
b
d


top predators
large sharks
smaller sharks
marlin
tuna
lancet fish
predators
squid
mackerel
filterers
lantern fish
amphipods
ocean sunfish
zooplankton
shrimp
copepods
pteropods
phytoplankton


(A)
Hubs forming a rich club
(B)
Shortest path
Long path
(C)
Triangular connections around node i
(D)
Modules
(E)
(F)


BOSTON
SACRAMENTO
OSAKA


Metallurgy
Fossil Fuel Mining & Utilities
Automobile
Machinery & Other Manufacturing
Real Estate, Finance/Insurance Health Care
Textiles
Plastics
Agriculture & Food
Link size and color (2002 billion $)
Node size (2002 billion $)
Node color
Agriculture
Mining
Utilities
Construction
Manufacturing
Finance/Insurance
Health Care
Real Estate
Wholesale/Retail Trade
Other Services
Government

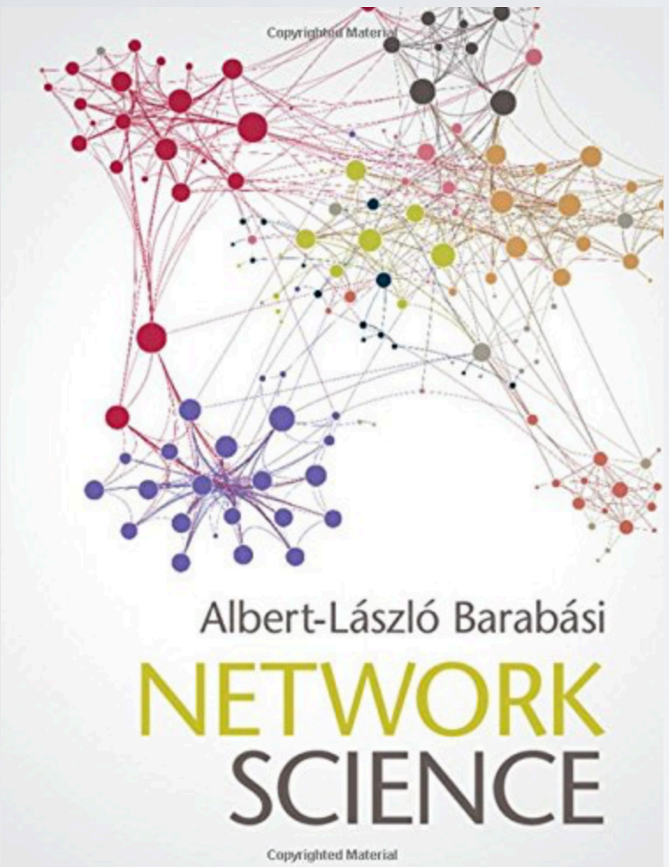
NETWORKS

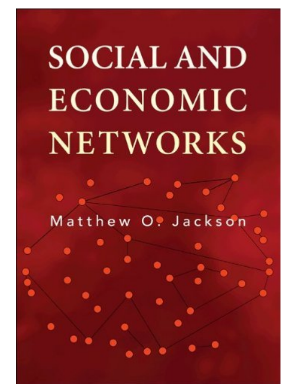
- Online social networks, e.g., Facebook, Twitter…
  - Nodes: accounts
  - Edges: relations (friend/follow) or interactions (wall post, like, retweet, mentions, etc.)

- Cryptocurrency
  - Nodes: addresses or *actors* (wallet ? Set of addresses ?)
  - Edges: transactions

NETWORK ANALYSIS

REFERENCES





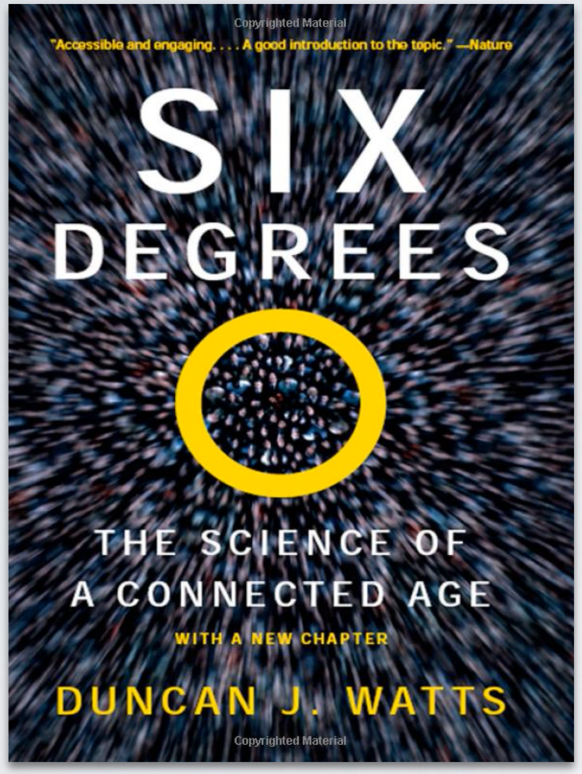
M. O. Jackson, Social and Economic Networks (Princeton University Press, 2010).

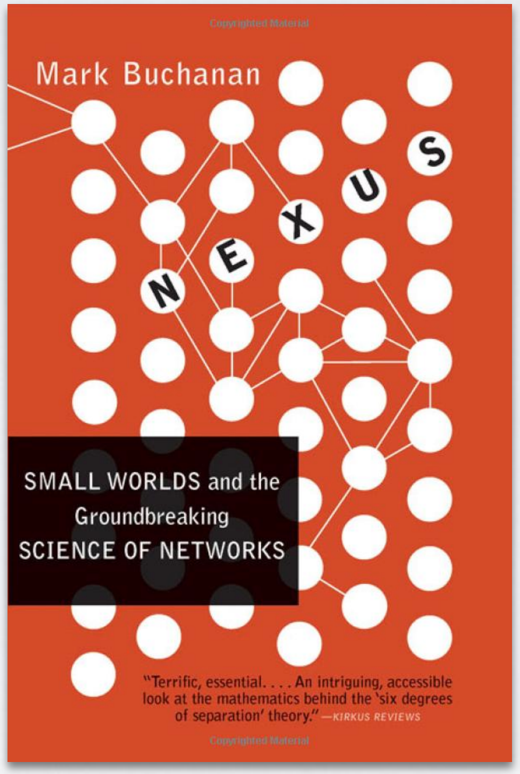
http://networksciencebook.com

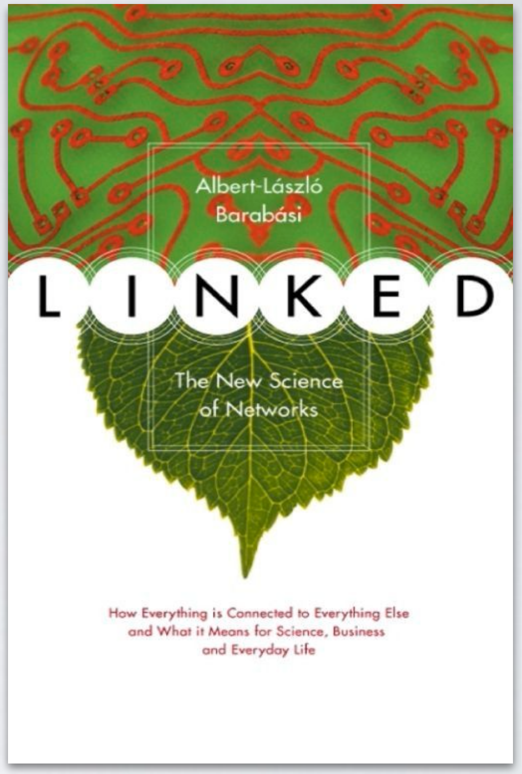
Google: "network science finance"
=> I'm not an expert in
economic networks :)

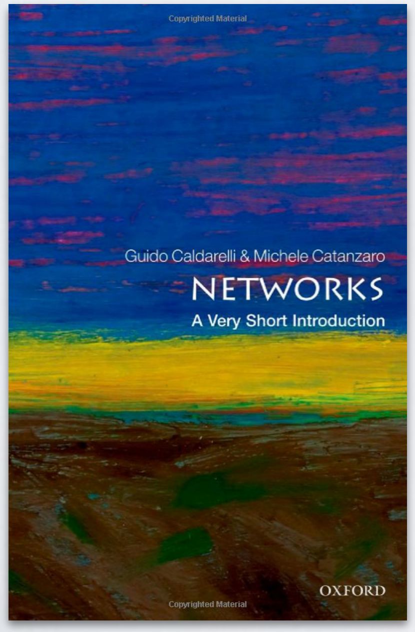
REFERENCES

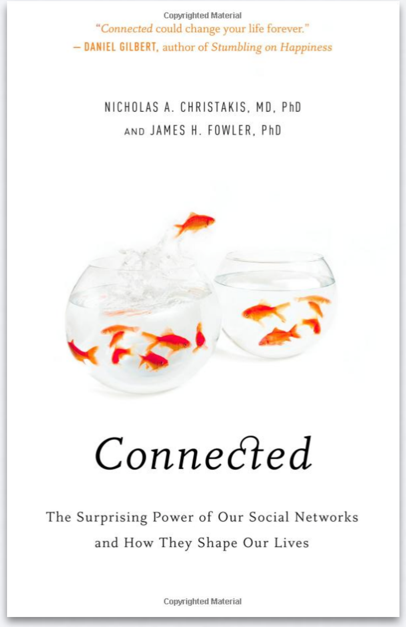
Pop-science books

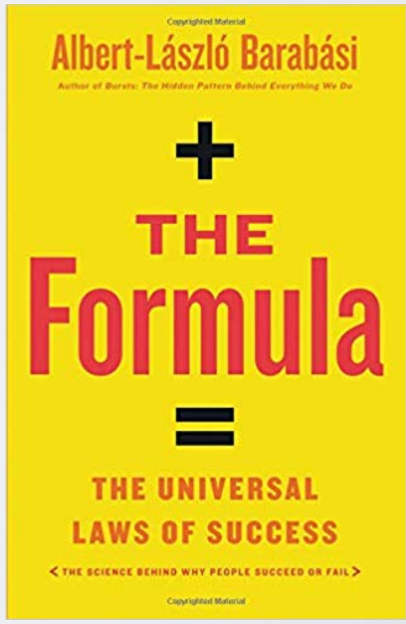




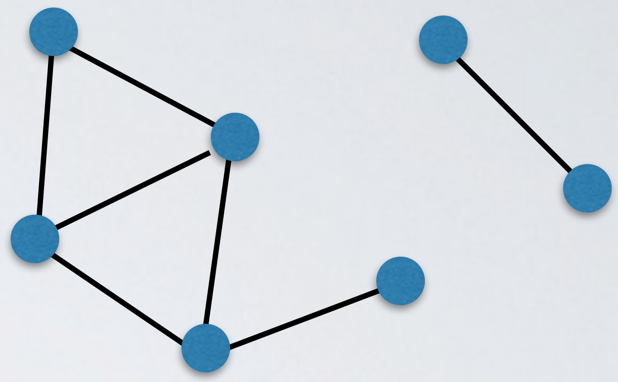








GRAPHS & NETWORKS



Networks often refers to real systems
- www,
- social network
- metabolic network.
- <u>Language: (Network, node, link)</u>

Vertex	Edge
person	friendship
neuron	synapse
Website	hyperlink
company	ownership
gene	regulation

Graph is the mathematical representation of a network
- <u>Language: (Graph, vertex, edge)</u>

In most cases we will use the two terms interchangeably.

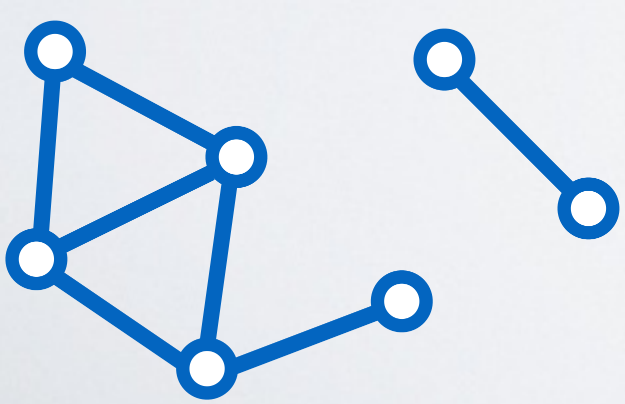
Types of Networks

Undirected networks

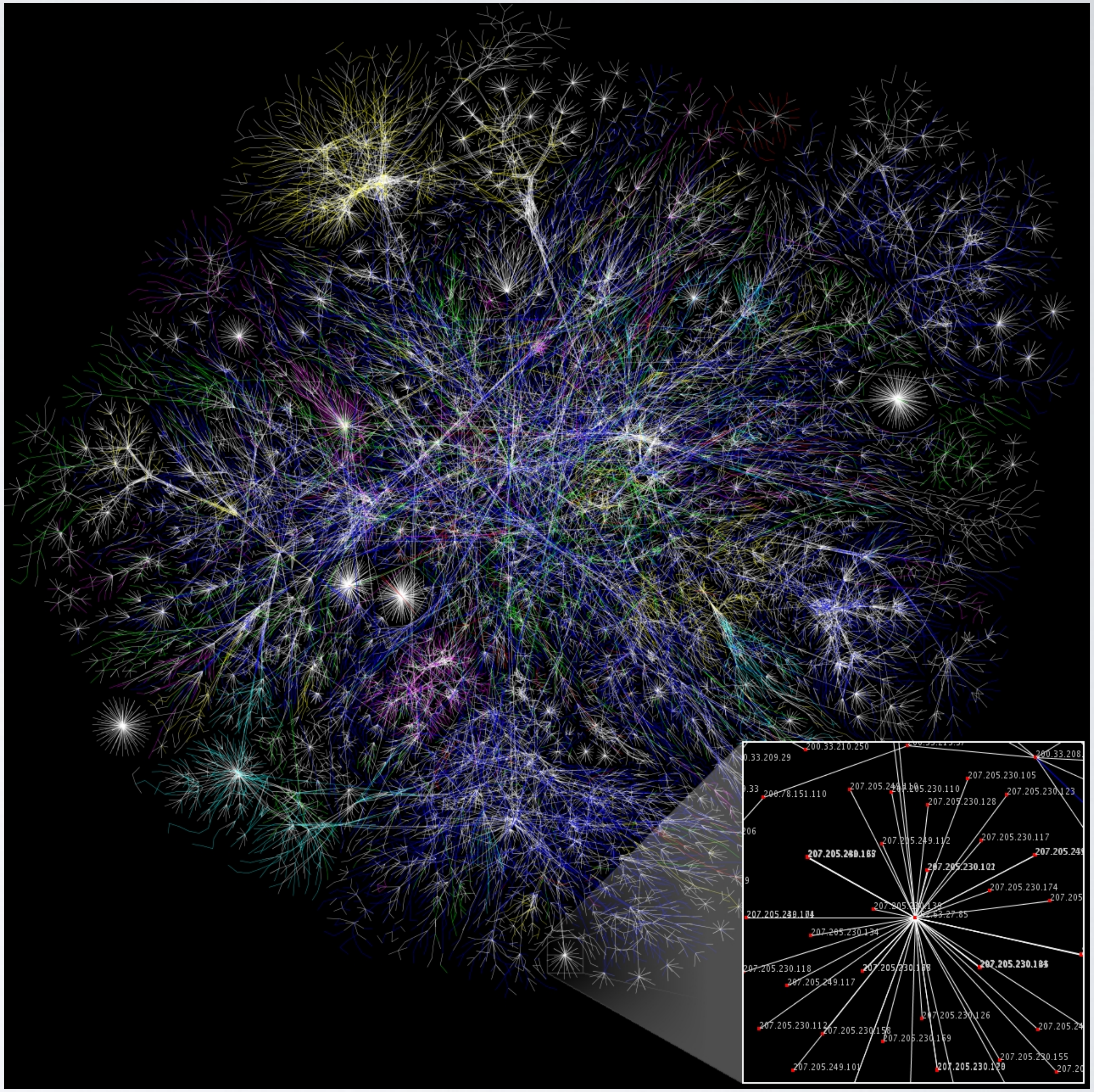
$G=(V, E)$

$(u,v) \in E \equiv (v,u) \in E$

- The directions of edges do not matter
- Interactions are possible between connected entities in both directions



Opte project



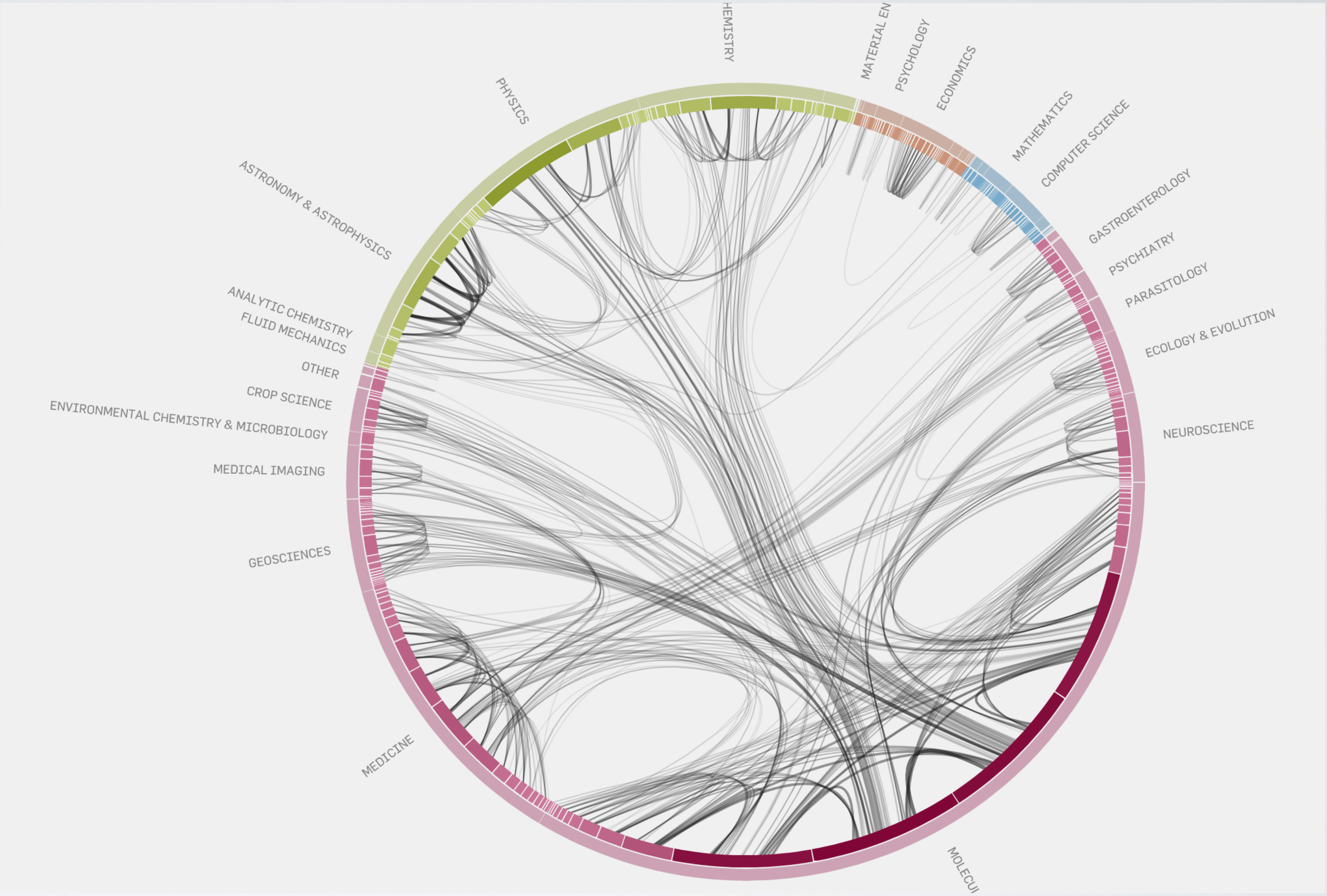
The Internet: Nodes - routers, Links - physical wires

Directed networks

Moritz Stefaner, eigenfactor.com

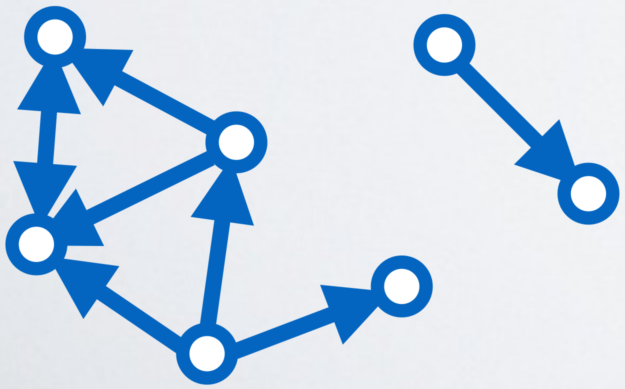
$G=(V, E)$

$(u,v) \in E \not\equiv (v,u) \in E$



Citation network: Nodes - publications, Links - references

- The directions of edges matter
- Interactions are possible between connected entities only in specified directions

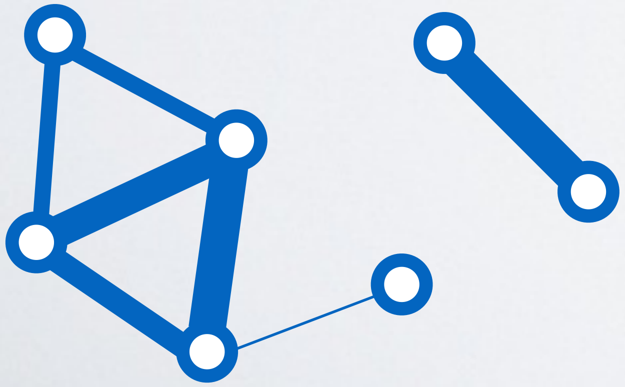


Weighted networks

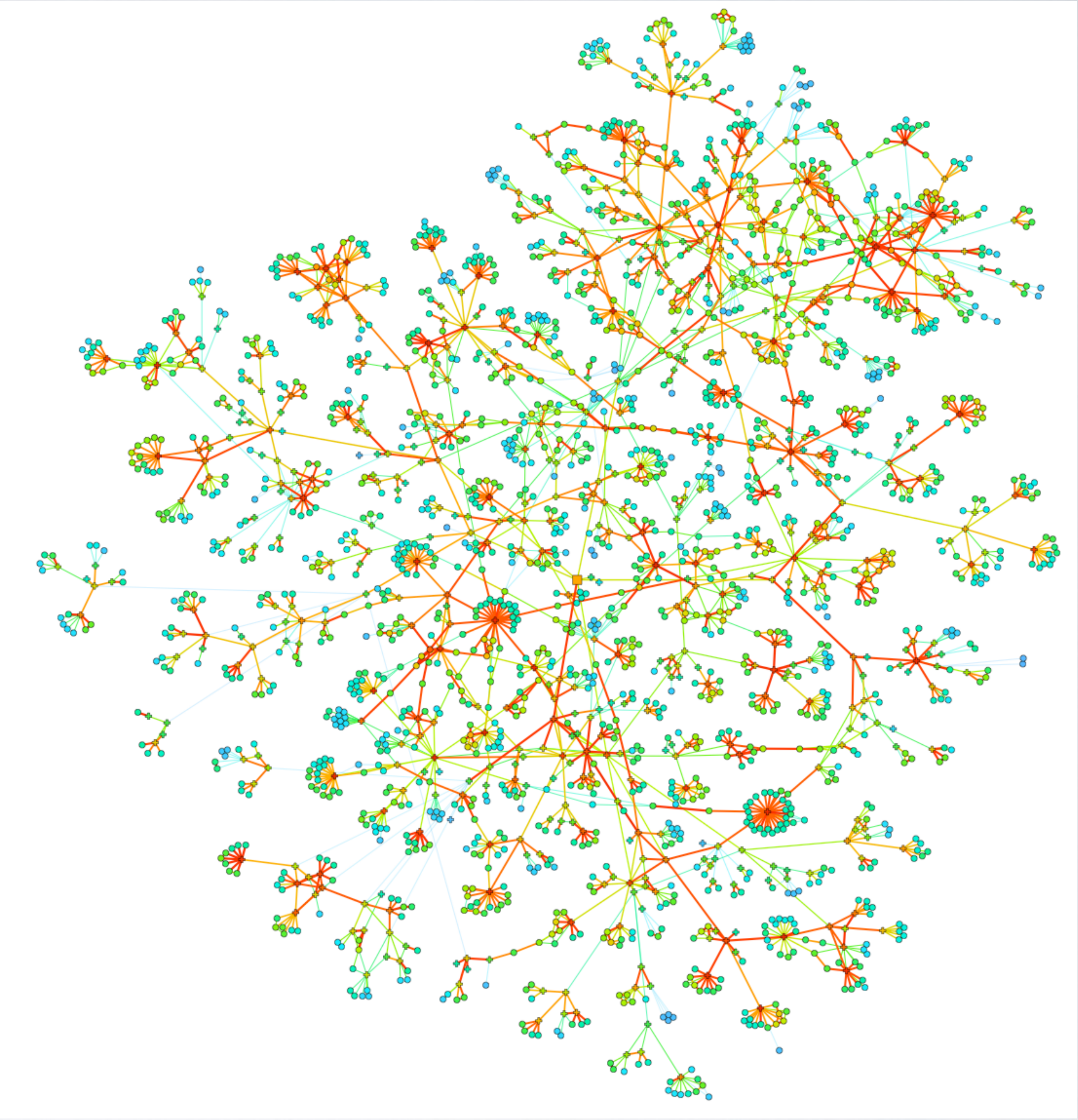
$G=(V, E, w)$

$w: (u,v) \in E \Rightarrow R$

- Strength of interactions are assigned by the weight of links



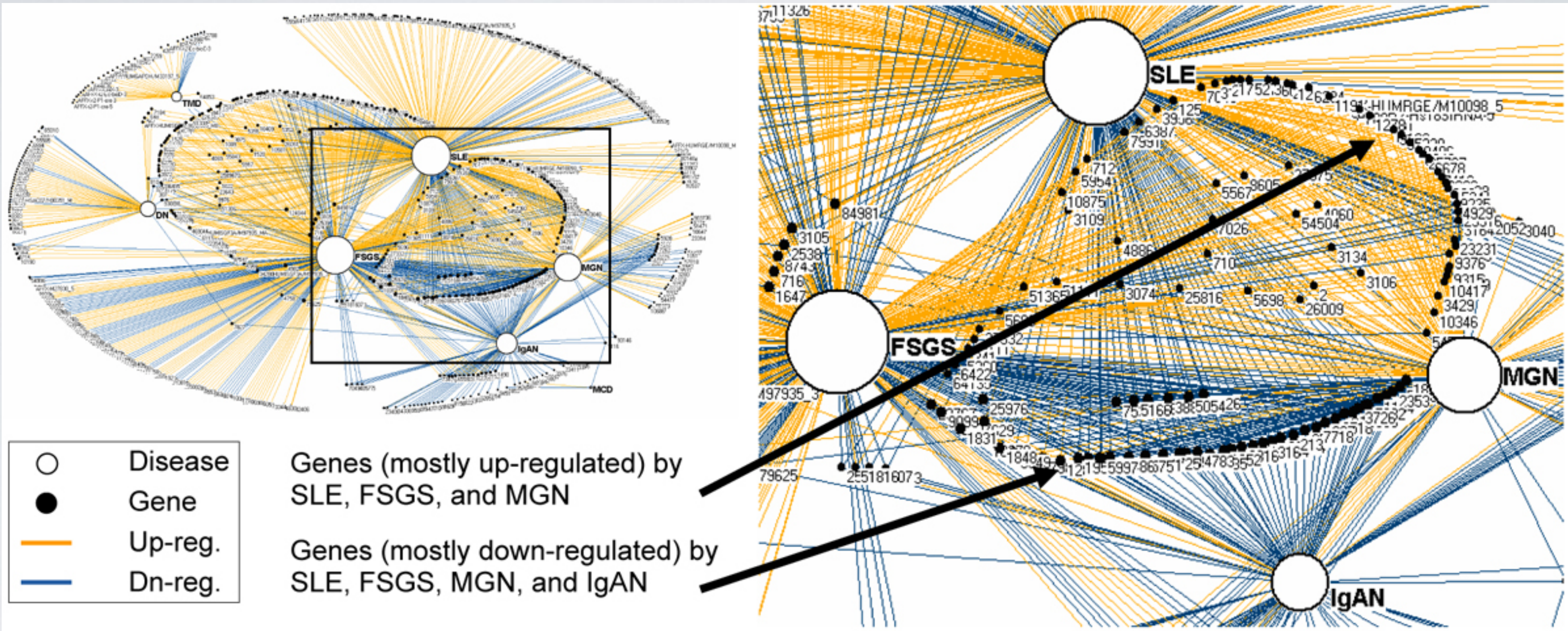
Onnela et.al. New Journal of Physics 9, 179 (2007).



Social interaction network: Nodes - individuals

Links - social interactions

Bipartite network

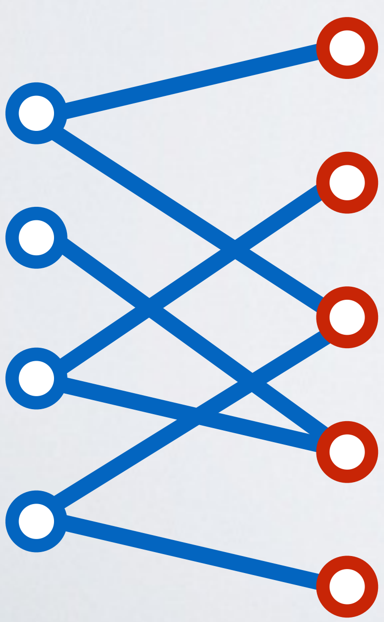


Bhavnani et.al. BMC Bioinformatics 2009, **10**(Suppl 9):S3

Gene-desease network:

Nodes - Desease (7)&Genes (747)

Links - gene-desease relationship



$$G=(U, V, E)$$

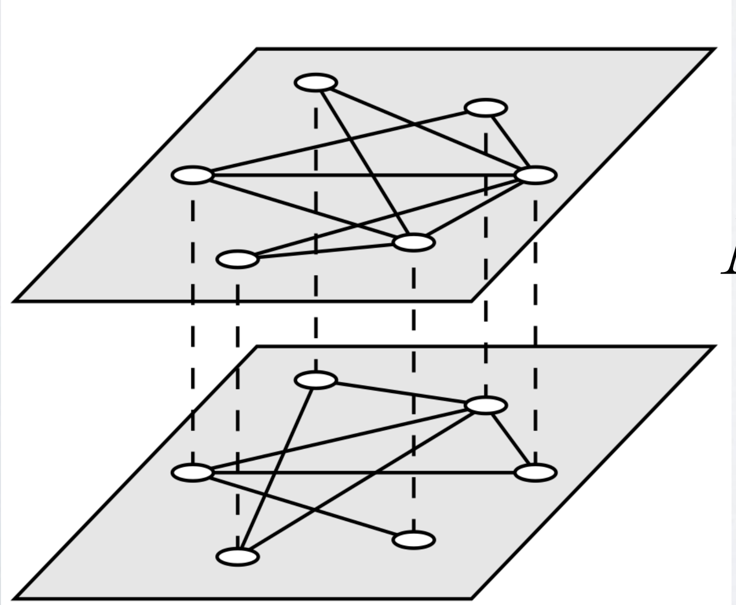
$$U \cap V = \varnothing$$

$$\forall (u,v) \in E,\ u \in U \text{ and } v \in V$$

Multiplex and multilayer networks

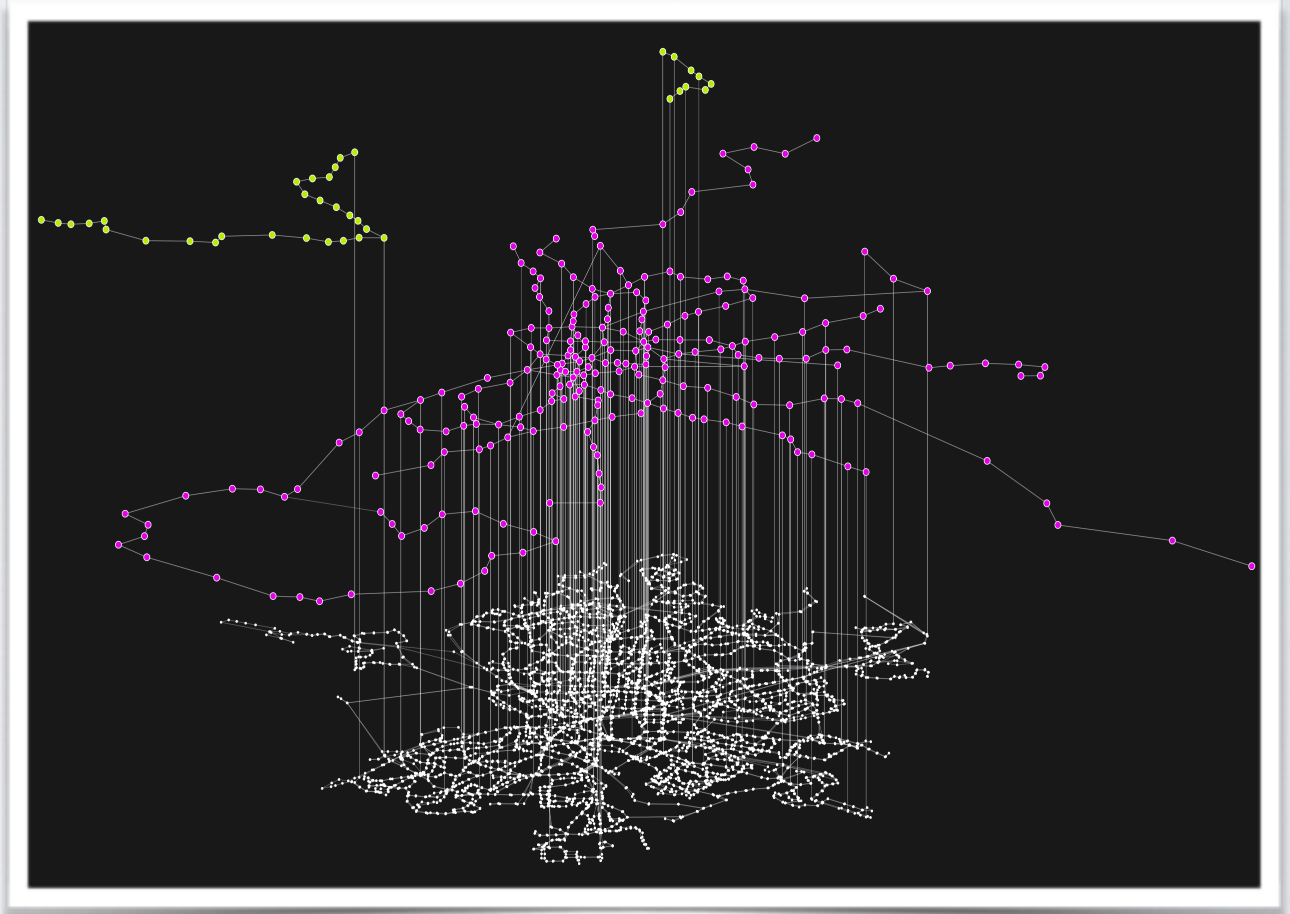
$G=(V, E_i),\ i=1\ldots M$

- Nodes can be present in multiple networks simultaneously
- These networks are connected (can influence each other) via the common nodes



$M=2$

Gomes et.al. Phys. Rev. Lett. 110, 028701 (2013)



[Mendez-Bermudez et al. 2017]

Temporal and evolving networks

$G=(V, E_t), (u,v,t,d) \in E_t$

t - time of interaction (u,v)
d - duration of interaction (u,v,t)

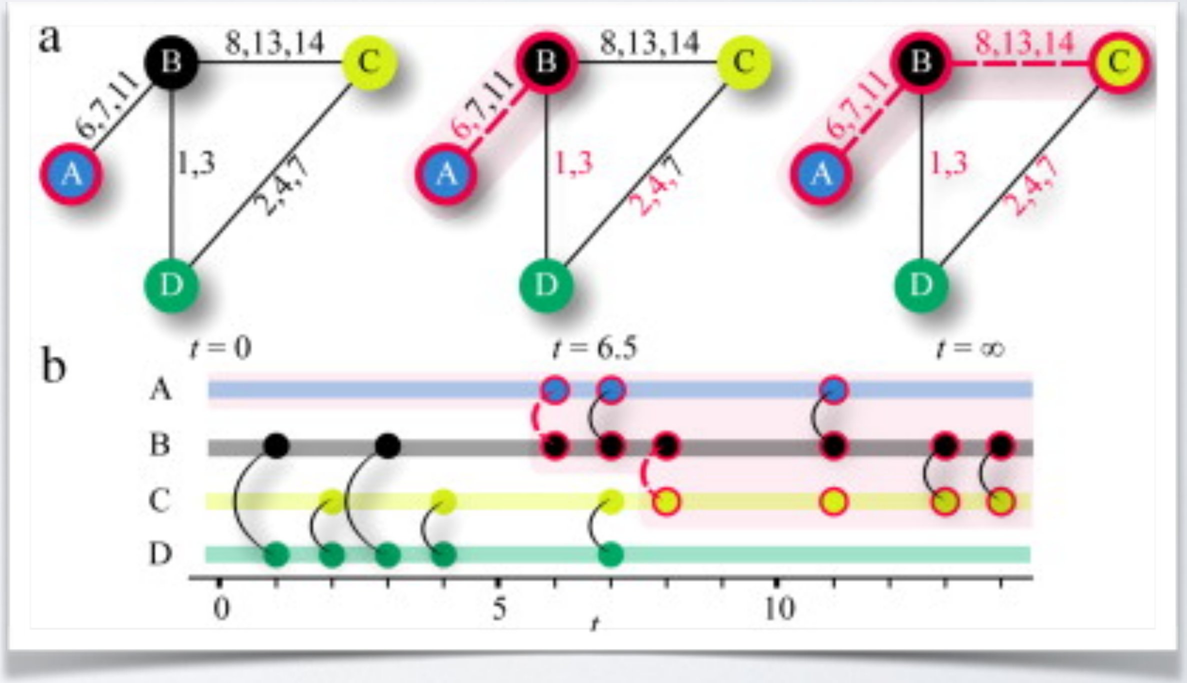
- Temporal links encode time varying interactions

$G=(V_{t'}, E_{t'})$

$v(t) \in V_{t'}$

$(u,v,t) \in E_{t'}$

- Dynamical nodes and links encode the evolution of the network



Mobile communication network
Nodes - individuals
Links - calls and SMS

COURSE OBJECTIVES

COURSE OBJECTIVES

- Theory:
  - Learn the basics of network science and network analysis, +some machine learning/data science concepts
- Practice:
  - Learn how to apply those concepts to graphs of small/medium size
- Project:
  - Apply what you learnt on a subset of the bitcoin transaction network

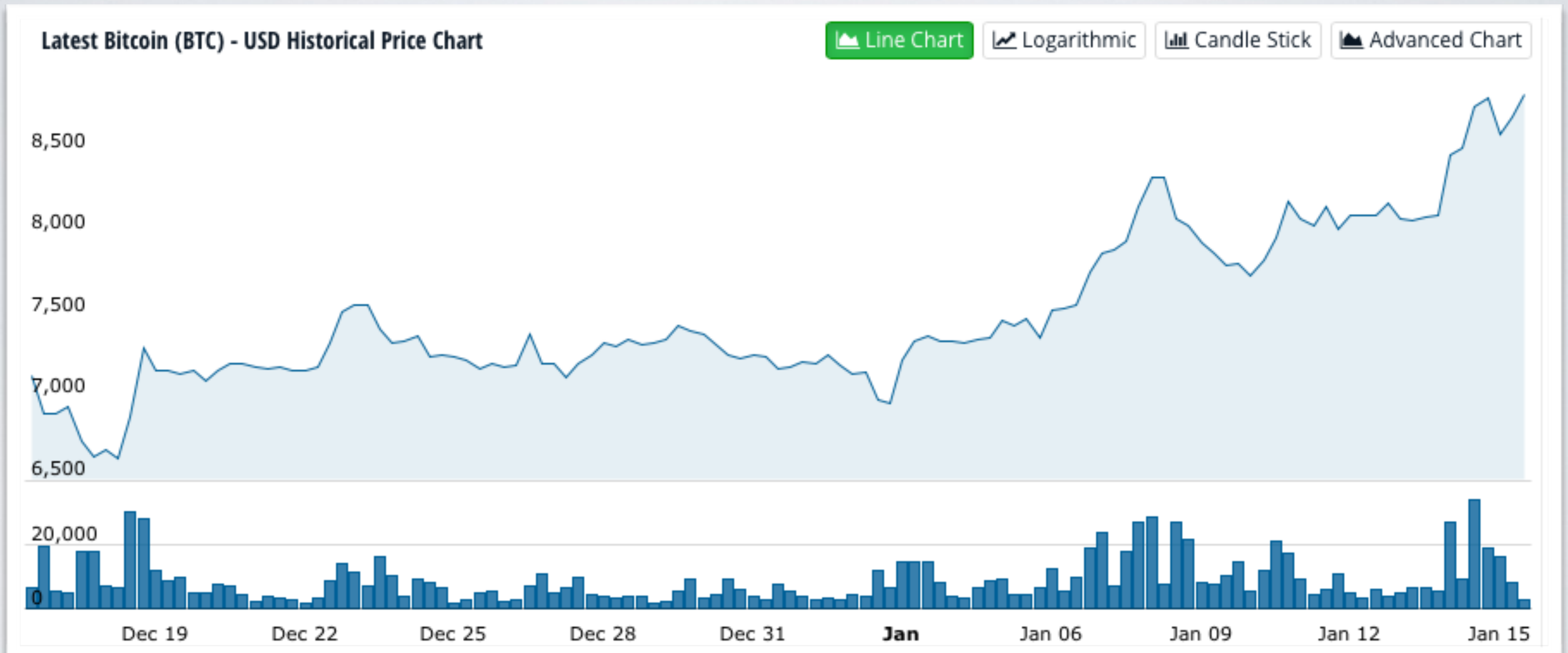
THE BITCOIN TRANSACTION NETWORK

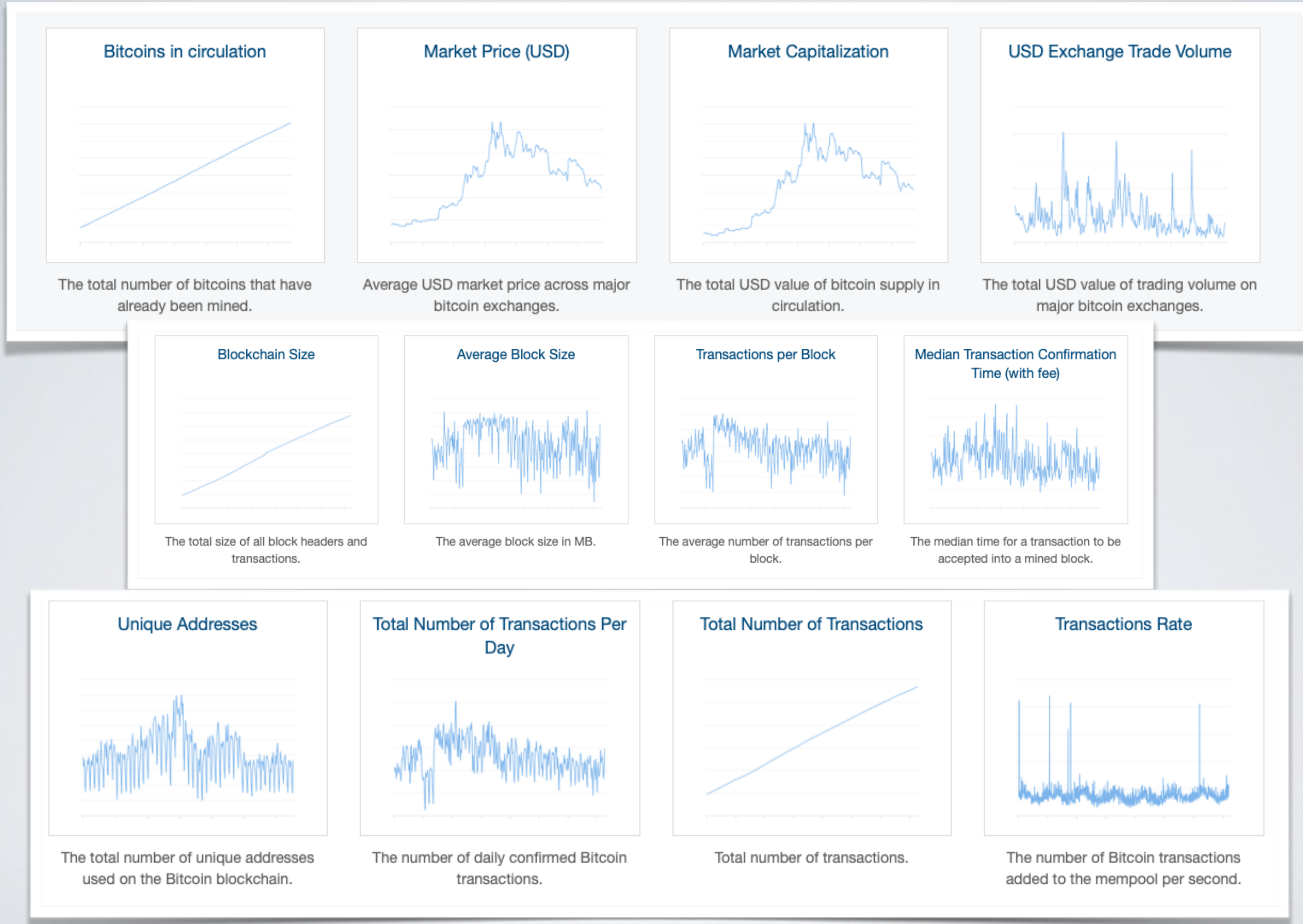
BITCOIN

- In this class, we are **not** interested in:
  - Cryptographic aspects
  - How the blockchain works
  - Governance of cryptocurrencies
  - Smart contracts
  - ICO
  - Macro-level analysis (transaction fee evolution, market price, etc.)

- What we are interested in:
  - Observing and understanding what is happening at the micro-level in one cryptocurrency (for this class, the largest one, Bitcoin) => **Look under the hood !**
  - How what is happening at the micro-level can be connected to what we observe at the macro-level (crisis, price fluctuation, macro-indicators…)

BITCOIN - MACRO LEVEL





https://www.blockchain.com/en/charts

BITCOIN - MACRO LEVEL

- This type of aggregated data is mostly identical to data you are used to in economy
- Can be studied with time series analysis (ARIMA, …)
- What is unique about Bitcoin:
  - We have all data about all transactions done using a given currency
  - We can use this information in relation with macro-level statistics
  - We can use it for new type of analysis

BITCOIN - DATA

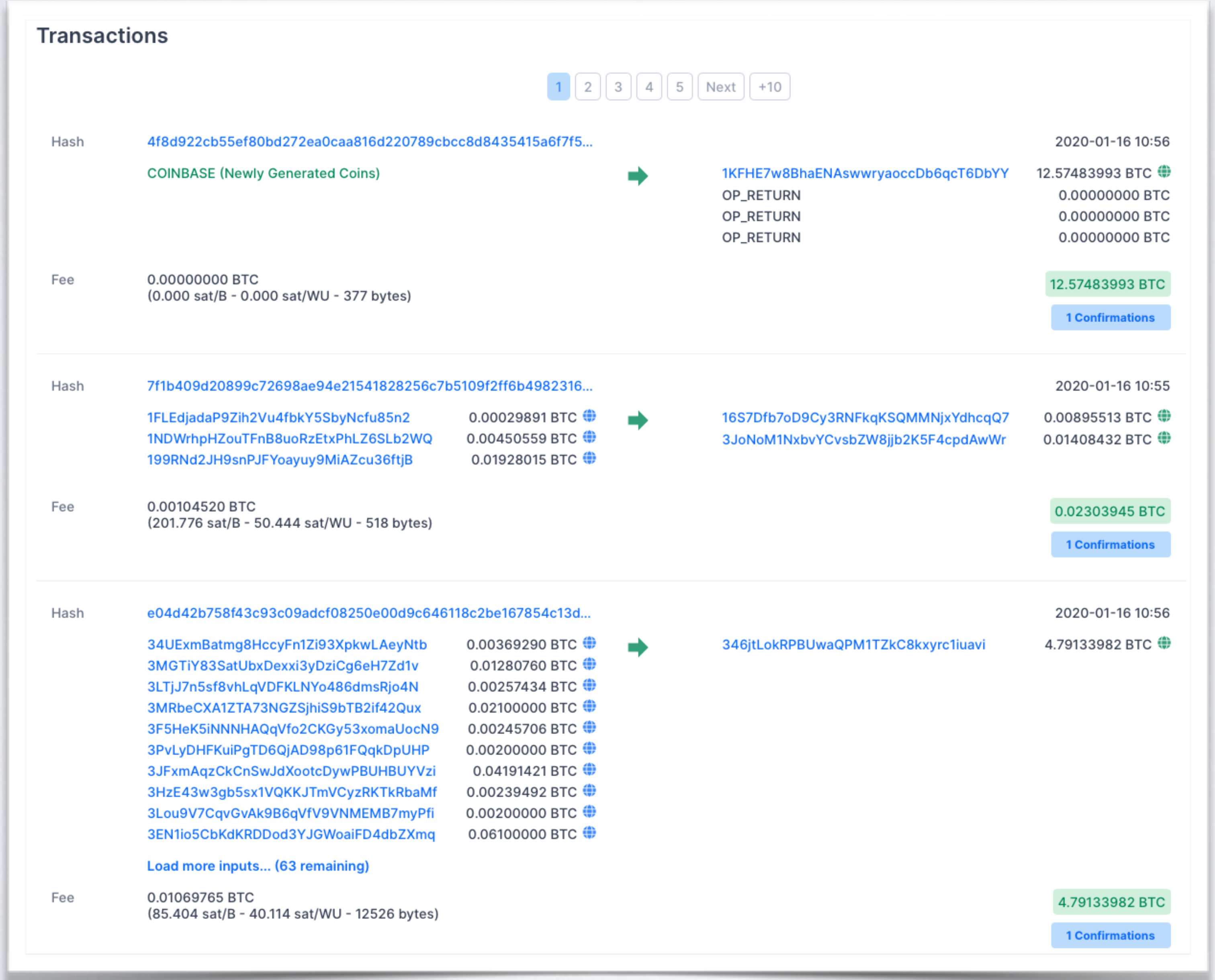
- The data we use: Content of the bitcoin blockchain
  - Seen as a simple list of transactions

Transaction	From	To	Value
t0	@1	@2	5
t1	@1	@3	2
…	…	…	…

- Bitcoin transactions are a little bit more complicated than that

BITCOIN - DATA

- You can explore it using tools such as a blockchain explorer
  - E.g.: https://www.blockchain.com/explorer



Hash	7f1b409d20899c72698ae94e21541828256c7b5109f2ff6b4982316...		2020-01-16 10:55
	1FLEdjadaP9Zih2Vu4fbkY5SbyNcfu85n2 0.00029891 BTC	16S7Dfb7oD9Cy3RNFkqKSQMMNjxYdhcqQ7	0.00895513 BTC
	1NDWrhpHZouTFnB8uoRzEtxPhLZ6SLb2WQ 0.00450559 BTC	3JoNoM1NxbvYCvsbZW8jjb2K5F4cpdAwWr	0.01408432 BTC
	199RNd2JH9snPJFYoayuy9MiAZcu36ftjB 0.01928015 BTC		
Fee	0.00104520 BTC (201.776 sat/B - 50.444 sat/WU - 518 bytes)		0.02303945 BTC 1 Confirmations

Hash	e04d42b758f43c93c09adcf08250e00d9c646118c2be167854c13d...		2020-01-16 10:56
	34UExmBatmg8HccyFn1Zi93XpkwLAeyNtb 0.00369290 BTC	346jtLokRPBUwaQPM1TZkC8kxyrc1iuavi	4.79133982 BTC
	3MGTiY83SatUbxDexxi3yDziCg6eH7Zd1v 0.01280760 BTC		
	3LTjJ7n5sf8vhLqVDFKLNYo486dmsRjo4N 0.00257434 BTC		
	3MRbeCXA1ZTA73NGZSjhiS9bTB2if42Qux 0.02100000 BTC		
	3F5HeK5iNNNHAQqVfo2CKGy53xomaUocN9 0.00245706 BTC		
	3PvLyDHFKuiPgTD6QjAD98p61FQqkDpUHP 0.00200000 BTC		
	3JFxmAqzCkCnSwJdXootcDywPBUHBUYVzi 0.04191421 BTC		
	3HzE43w3gb5sx1VQKKJTmVCyzRKTkRbaMf 0.00239492 BTC		
	3Lou9V7CqvGvAk9B6qVfV9VNMEMB7myPfi 0.00200000 BTC		
	3EN1io5CbKdKRDDod3YJGWoaiFD4dbZXmq 0.06100000 BTC		
	Load more inputs... (63 remaining)		
Fee	0.01069765 BTC (85.404 sat/B - 40.114 sat/WU - 12526 bytes)		4.79133982 BTC 1 Confirmations

UNDERSTANDING BITCOIN TRANSACTIONS

- Transactions are between *m* "inputs" and *n* "outputs"
- Each *input* (resp. *output)* is a pair (value,bitcoin address)
- *inputs* are necessarily *outputs* of previous transactions
  - Unlocked by the private key of the payer

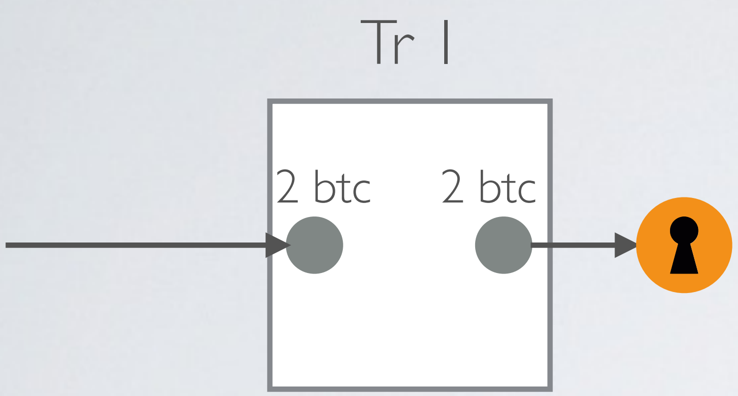
UNDERSTANDING BITCOIN TRANSACTIONS

- A user possess a **private key**

- A user can generate **public keys** (bitcoin adresses)
  - Instantaneously
  - At no cost
  - As often as wanted

- Public key ≈ lock that can be opened only by an associated private key

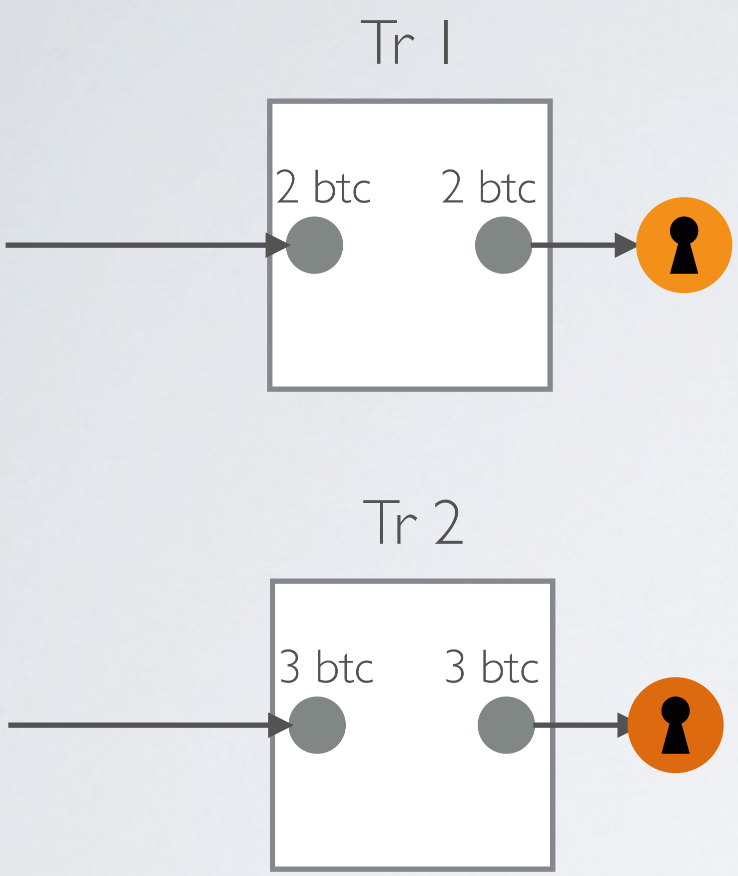
ILLUSTRATION



Public keys of user U1 :

1BusVkYQvbbGbSDZNo5DfhrFeQdgK1Y1VY

ILLUSTRATION

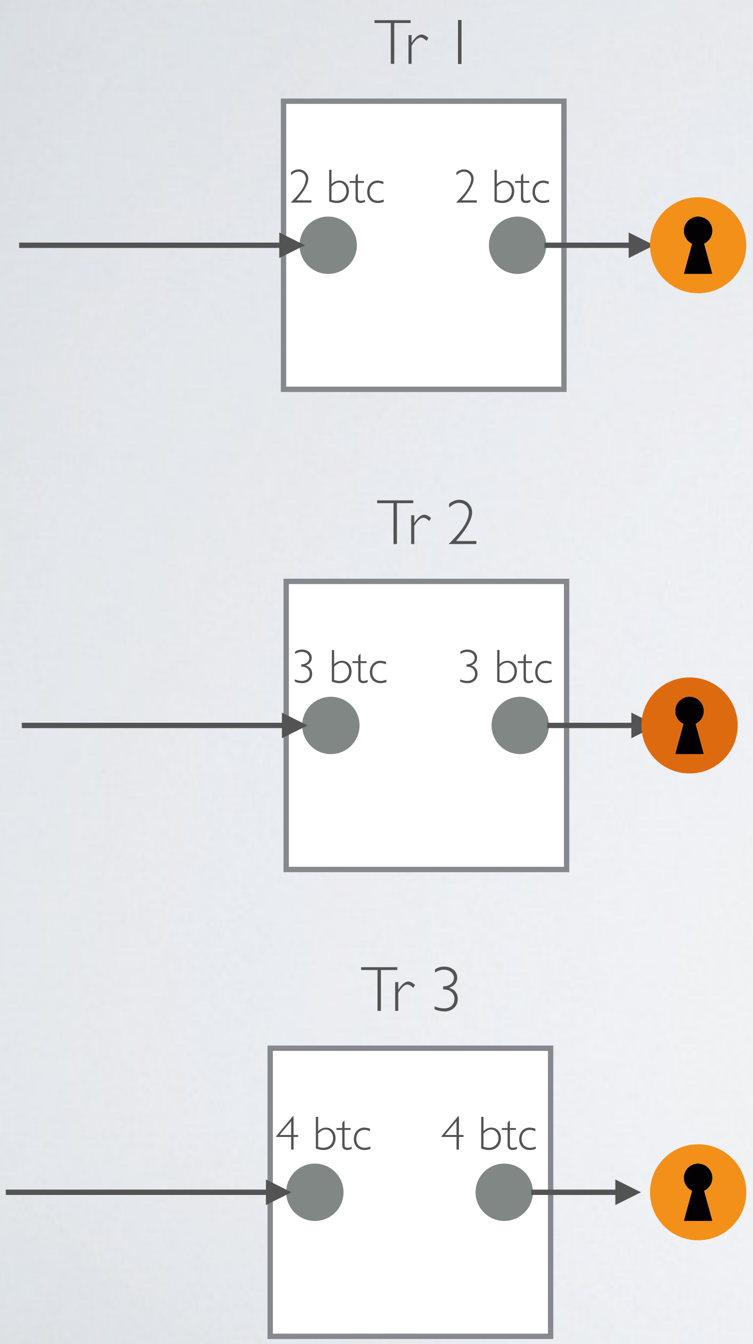


Public keys of user U1 :

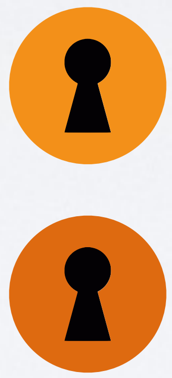
1BusVkYQvbbGbSDZNo5DfhrFeQdgK1Y1VY

1QFdbGkhiCDFF45mBHgzWUdiqv55NJbd4u

ILLUSTRATION



Public keys of user U1 :



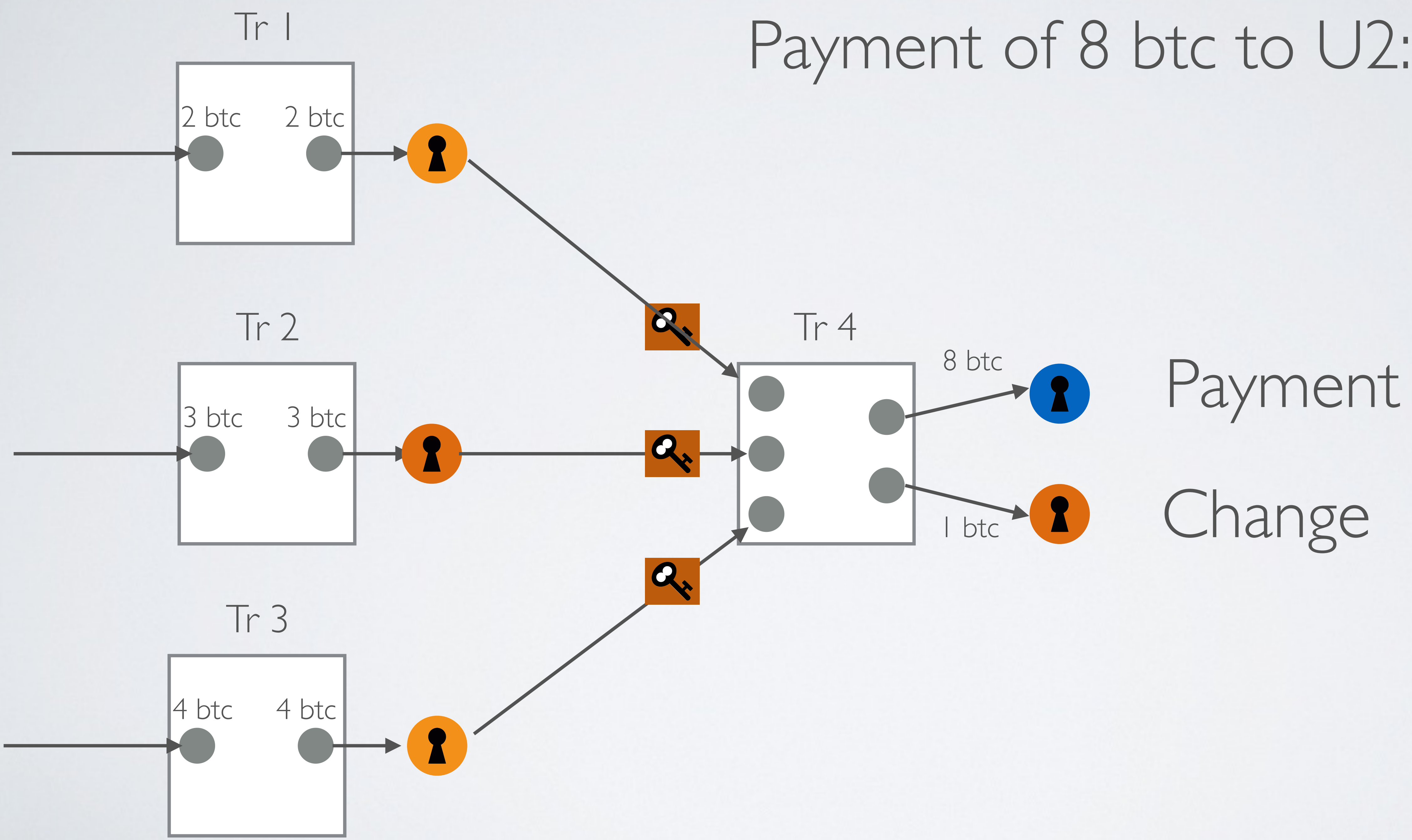
1BusVkYQvbbGbSDZNo5DfhrFeQdgK1Y1VY

1QFdbGkhiCDFF45mBHgzWUdiqv55NJbd4u

“Wallet” of U1:

- 9 btc
- Divided in 3 “output”
- Locked by 2 different public keys

ILLUSTRATION



ADDRESS NETWORK

- First network, node=Address
  - Naive approach
  - One address ≠ one user!
- Node: bitcoin address (public key)
- Edge: input addresses to output addresses.
- Problem: most transactions have several inputs, several outputs
  - Values ?

ADDRESS NETWORK

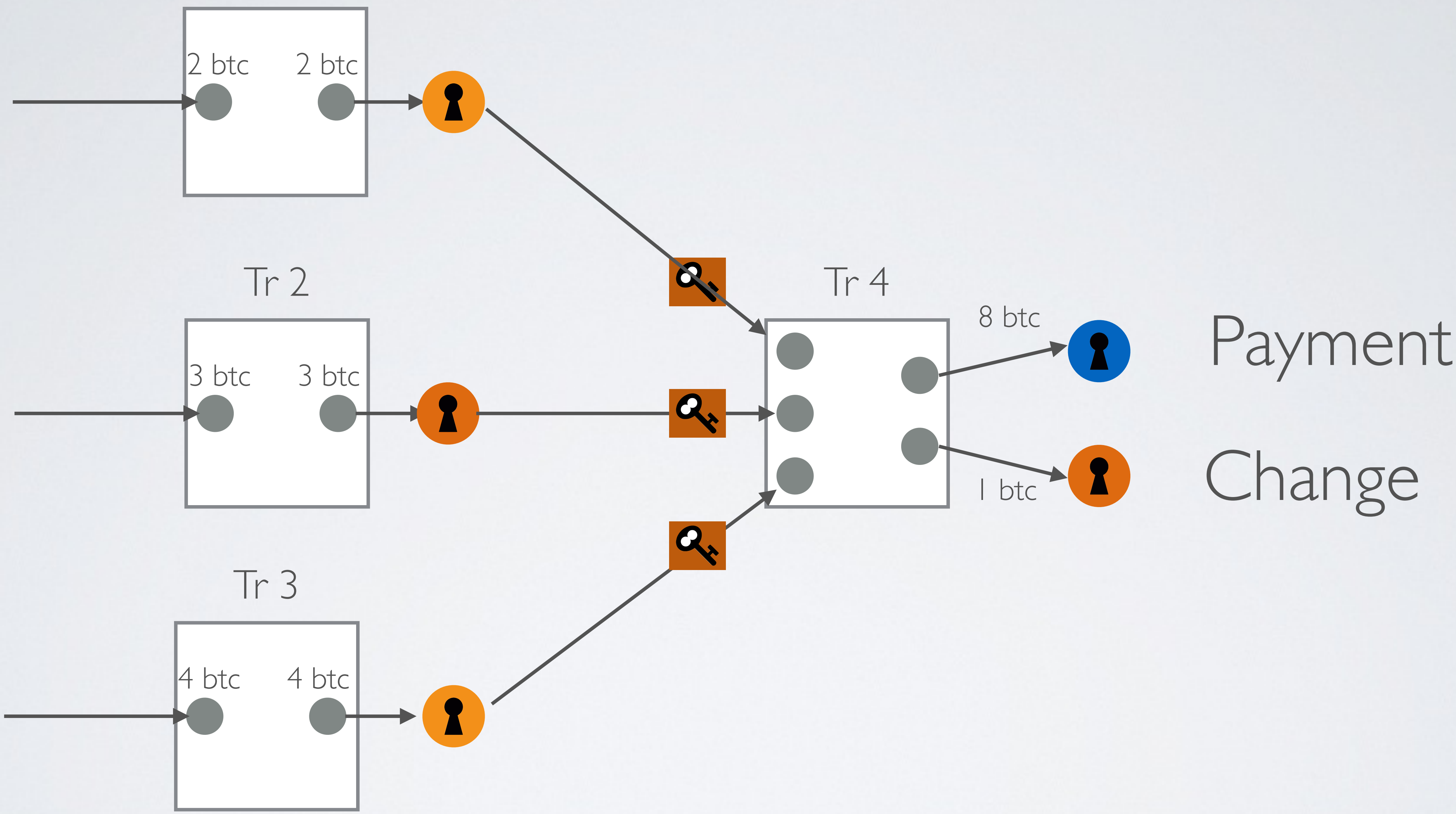
- # Transactions: 490 441
- # Transaction outputs: 1 210 004 (avg. 2,46)
- # Transaction inputs 1 211 790 (avg. 2.47)
- # Addresses: 933 645
- # @->@ Edges: 3 014 350

- Very big, hard to interpret

ACTOR NETWORK

- Transactions between "actors" of the bitcoin ecosystem
  - Individuals with their own private key (e.g., using BRD, Atomic Wallet, etc.)
  - Companies/organisations with their own private key
  - Exchanges (e.g., Binance, CoinBase, etc.)
  - Mining Pool
  - etc.
- An actor has **one** private key, but can have **many** public keys/addresses
- How to retrieve addresses belonging to the same actor?

ACTOR NETWORK



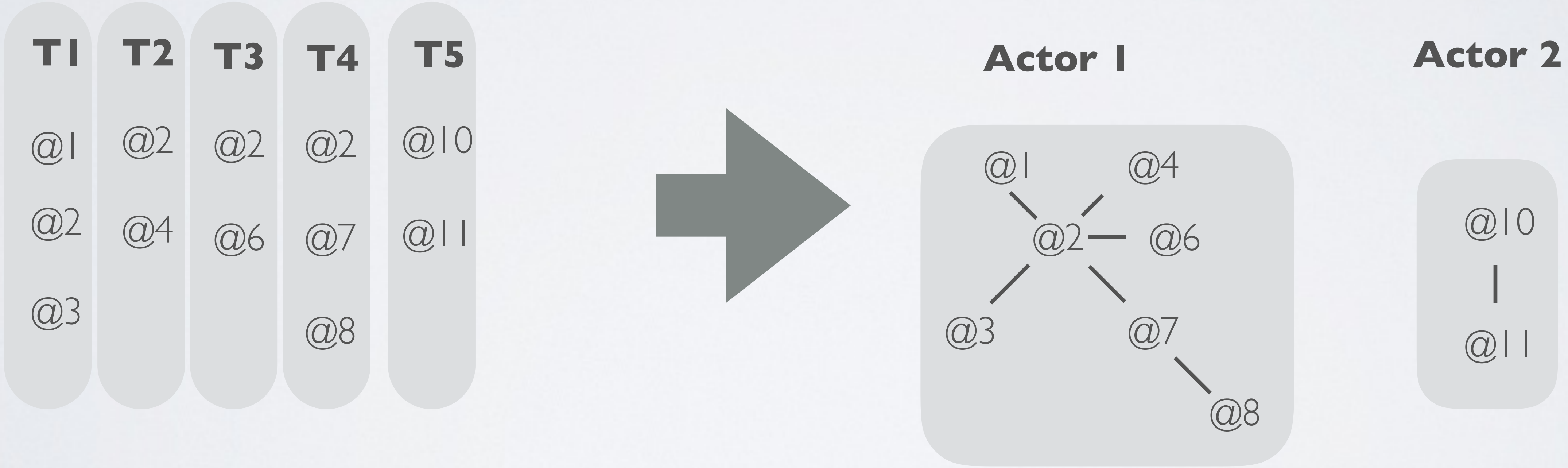
and are inputs of the same transaction

=> same actor

ACTOR NETWORK

- Actor identification: find all addresses of a same user
  - Currently a research question…
- Heuristics (input):
  - All addresses in input of a same transaction belongs to the same person

ACTOR NETWORK

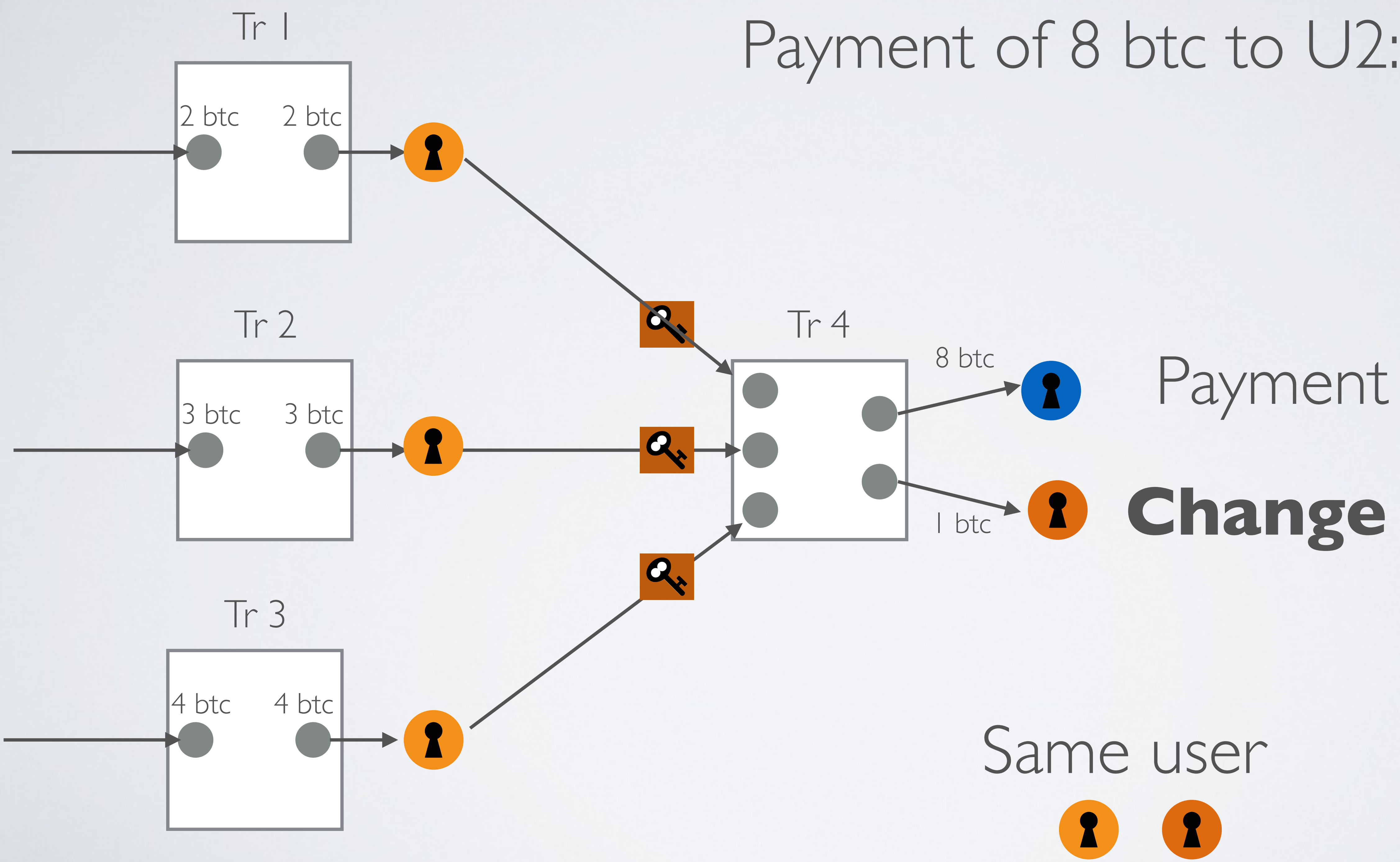


ACTOR NETWORK

- Actor identification: find all addresses of a same user
  - Currently a research question…
- Heuristics (input):
  - All addresses in input of a same transaction belongs to the same person
- Heuristics (output):
  - One of the addresses in output is probably a **change address**, thus an address of the same user as the one in input
  - But which one ?

ACTOR NETWORK

Payment of 8 btc to U2:



Same user

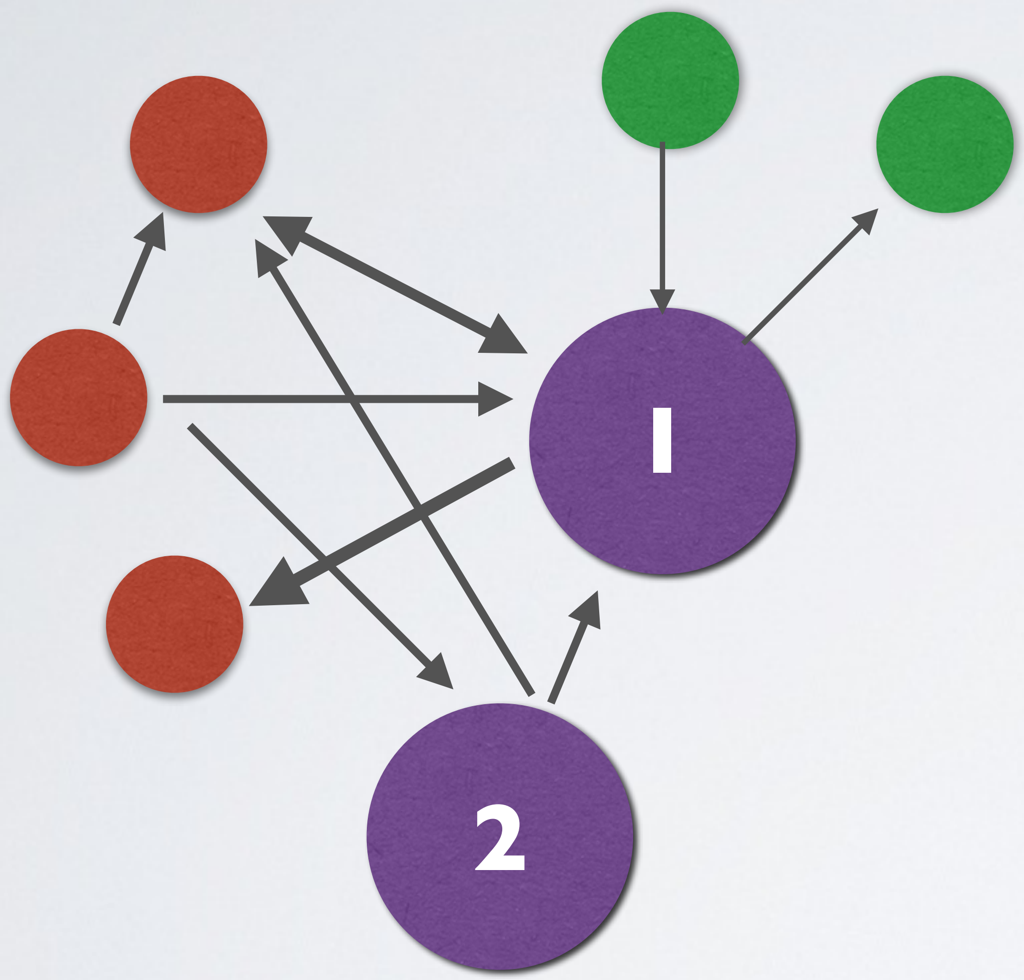
ACTOR NETWORK

- Heuristics (output):
  - One of the addresses in output is probably a **change address**, thus an address of the same user as the one in input
  - But which one ?
    - Lower value ?
    - Value with the same decimal as input?
    - Learn which one using machine learning and examples ?
    - …
    - => A research question, not in the scope of this class.
  -

ACTOR NETWORK

- Group of addresses => Anonymous actor
  - Can we know who is this actor?
  - It is enough to identify *one* address
  - One transaction with a person/company => we know one of its addresses
  - On the internet, many company/individuals provide their addresses.
  - For some actors, we might infer their category
    - => Miners
    - => Large transactions profiles VS low transaction profiles
    - Has made transactions to identified money laundering services => suspicious
    - Machine learning => Automatically recognize profiles, identify similar actors, …
    - etc.

OBTAINED NETWORK

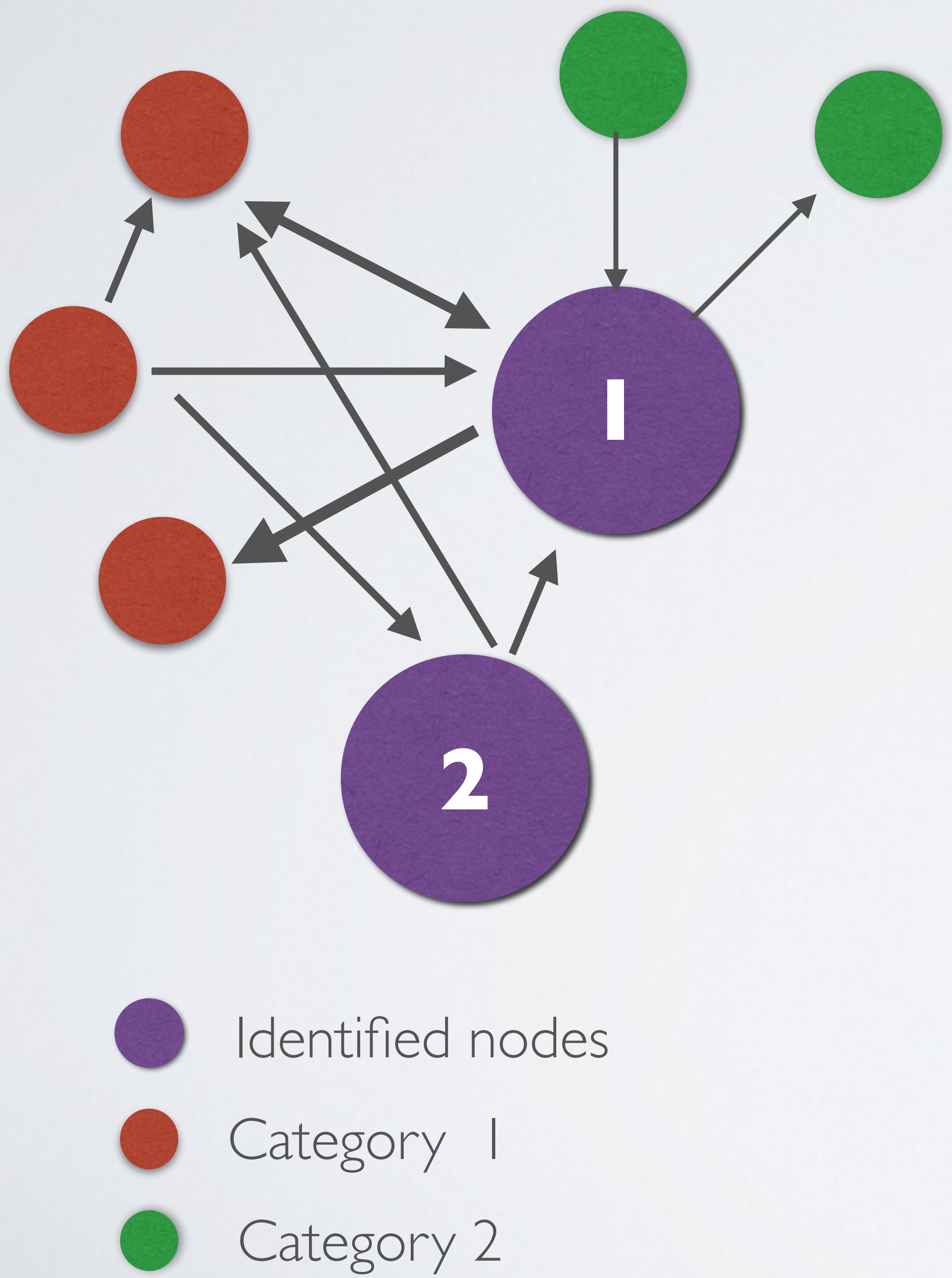


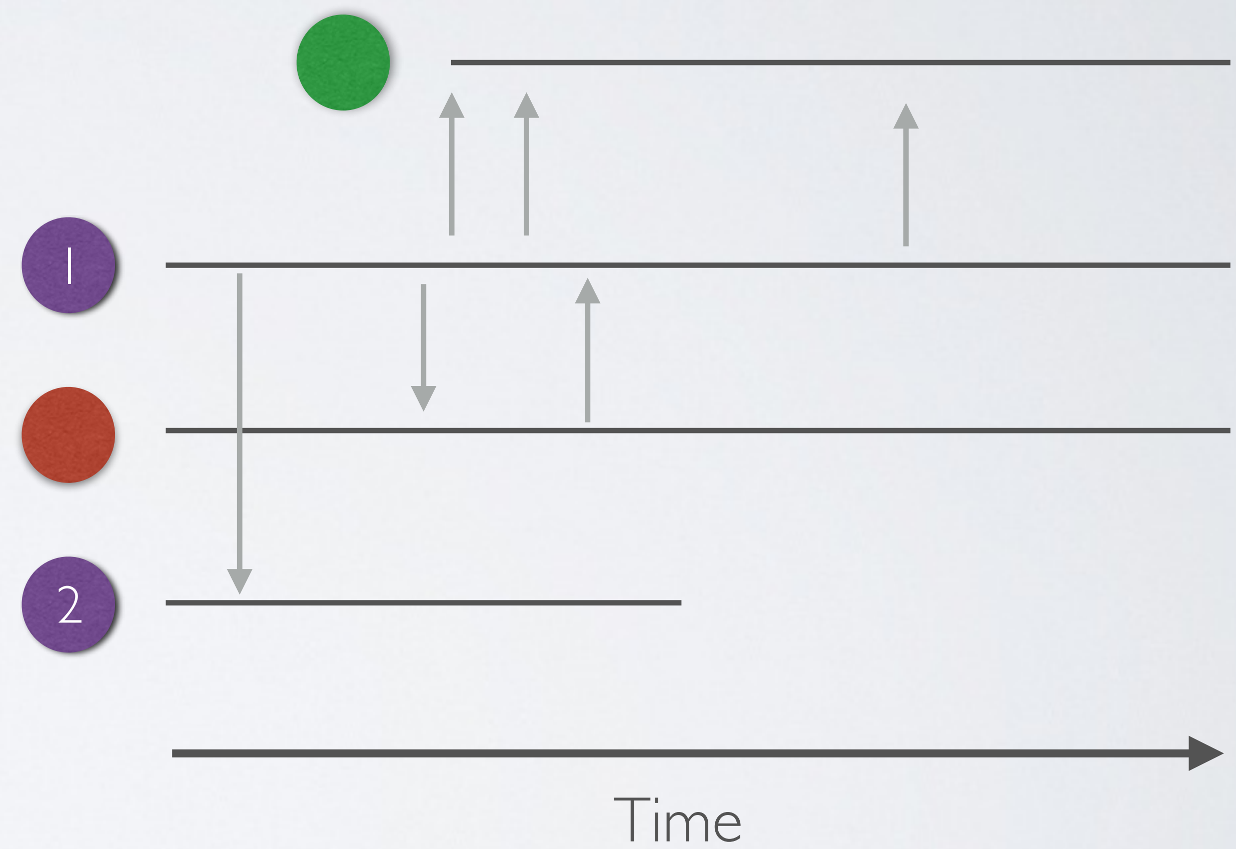
Identified nodes

Category 1

Category 2

OBTAINED NETWORK





ADDRESS NETWORK

- Example: 2 days (August 2&3 2016)
- Address network
  - # Transactions: 490 441
  - # Transaction outputs: 1 210 004 (avg. 2,46)
  - # Transaction inputs 1 211 790 (avg. 2.47)
  - # Addresses: 933 645
  - # @->@ Edges: 3 014 350
- Actor network
  - # Clusters: 456 012
  - Largest clusters sizes: 20 023, 19 381, 17 244
  - # Actor -> Actor Edges : 956 347

GRAPH DESCRIPTION

DESCRIPTION OF GRAPHS

- When confronted with a graph, how to describe it?
- How to compare graphs?
- What can we say about a graph?

SIZE

- A network is composed of nodes and edges.
- Size: How many nodes and edges ? (n & m)

	#nodes (n)	#edges (m)
Wikipedia HL	2M	30M
Twitter 2015	288M	60B
Facebook 2015	1.4B	400B
Brain c. Elegans	280	6393
Roads US	2M	2.7M
Airport traffic	3k	31k

DENSITY

Defined as:

Directed
$$D = \frac{|E|}{|V|\,(|V|-1)}$$

Undirected
$$D = \frac{2|E|}{|V|\,(|V|-1)}$$

Often more relevant: average degree (2|E| / |V|)

	#nodes	#edges	Density	avg. deg
Wikipedia	2M	30M	1.5×10^{-5}	30
Twitter 2015	288M	60B	1.4×10^{-6}	416
Facebook	1.4B	400B	4×10^{-9}	570
Brain c.	280	6393	0,16	46
Roads Calif.	2M	2.7M	6×10^{-7}	2,7
Airport	3k	31k	0,007	21

DENSITY

- It has been observed that: [Leskovec. 2006]
  - When graphs increase in size, the average degree increases
  - This increase is very slow
- Think of friends in a social network

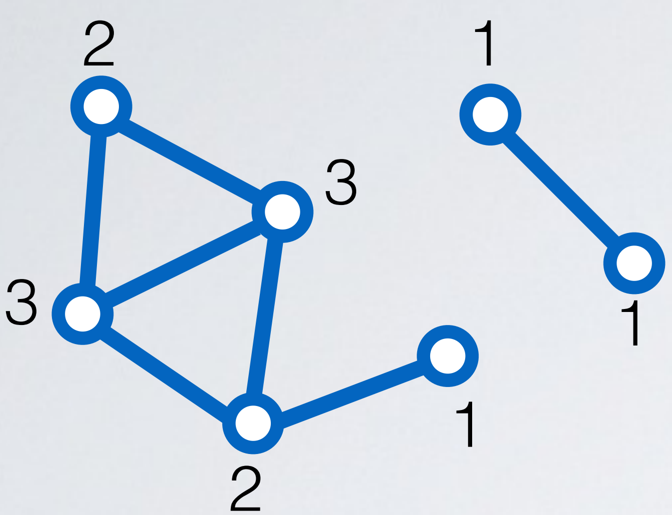
Node degree

Number of connections of a node

- Undirected network

$$k_i = A_{i1} + A_{i2} + ... + A_{iN} = \sum_j^N A_{ij}$$

$$m = \frac{\sum_i k_i}{2} \quad \text{where} \quad m = |E|$$

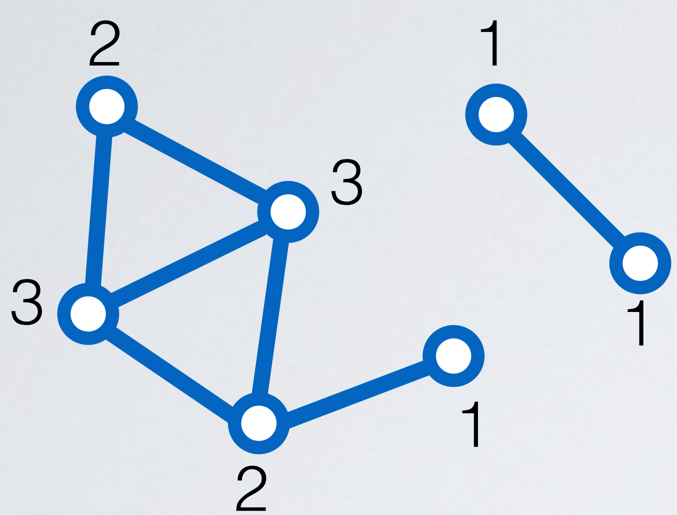


	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

Node degree

Number of connections of a node

- Undirected network



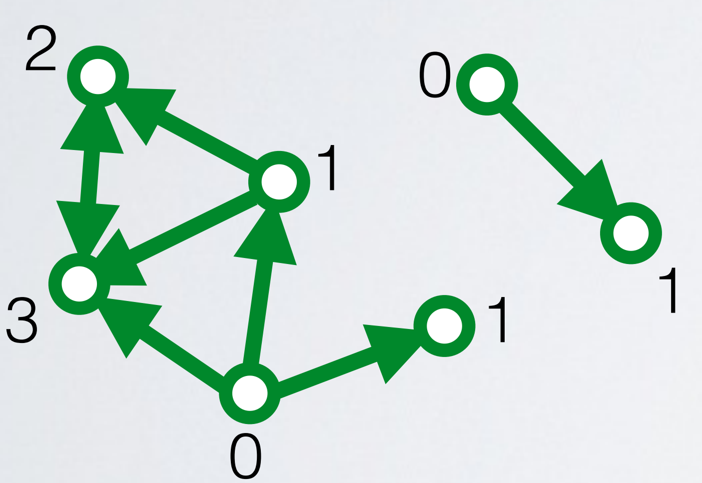
$$k_i = A_{i1} + A_{i2} + \ldots + A_{iN} = \sum_j^N A_{ij}$$

$$m = \frac{\sum_i k_i}{2} \quad \text{where} \quad m = |E|$$

mean degree $$\langle k \rangle = \frac{1}{N} \sum_i^N k_i$$

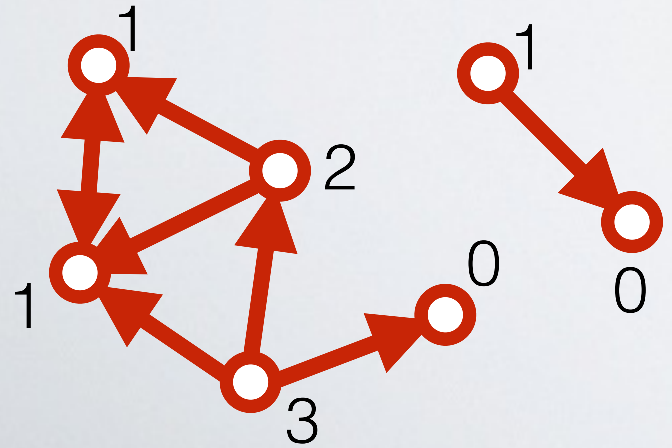
	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

- Directed network



In degree $$k_i^{in} = \sum_j^N A_{ij}$$

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0



Out degree $$k_j^{out} = \sum_i^N A_{ij}$$

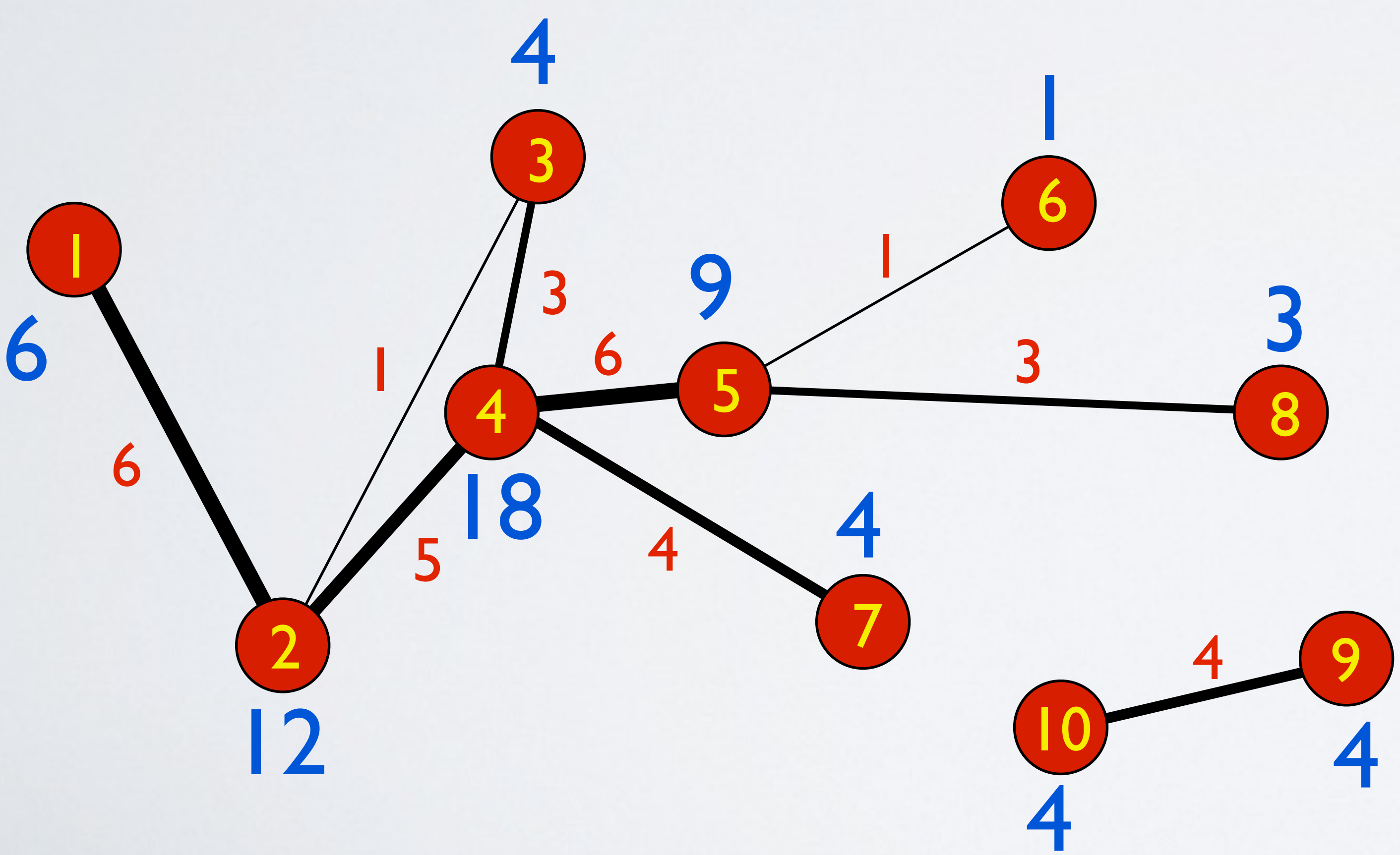
	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

Weighted degree: strength

- Weighted networks

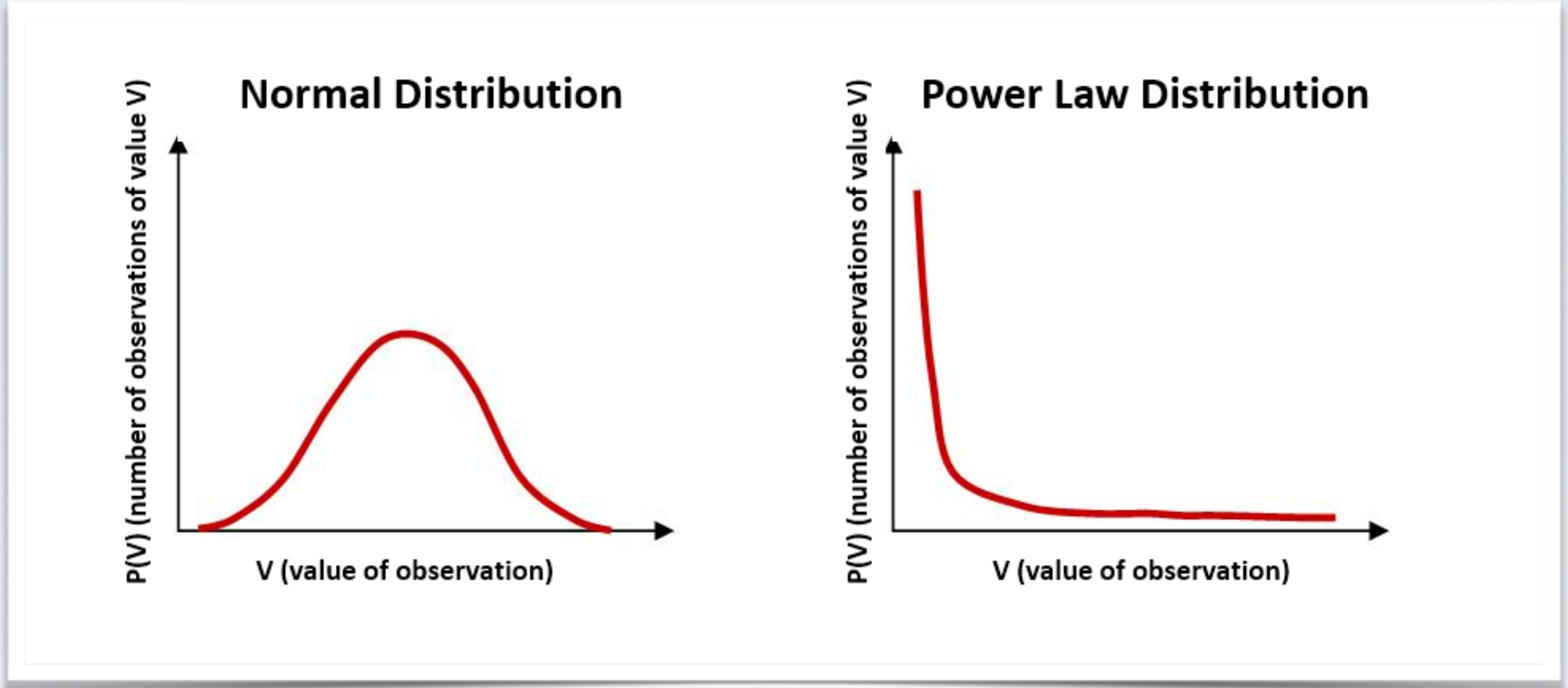
The sum of the weights of links connected to node i

$$s_i = w_{i1} + w_{i2} + \ldots + w_{iN} = \Sigma_j \, w_{ij}$$



	1	2	3	4	5	6	7	8	9	10
1	0	6	0	0	0	0	0	0	0	0
2	6	0	1	3	0	0	0	0	0	0
3	0	1	0	3	0	0	0	0	0	0
4	0	3	3	0	6	0	4	0	0	0
5	0	0	0	6	0	1	0	3	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	4	0	0	0	0	0	0
8	0	0	0	0	3	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	4
10	0	0	0	0	0	0	0	0	4	0

DEGREE DISTRIBUTION



PDF (Probability Distribution Function)

Sometimes with CDF (Cumulative Distribution Function)

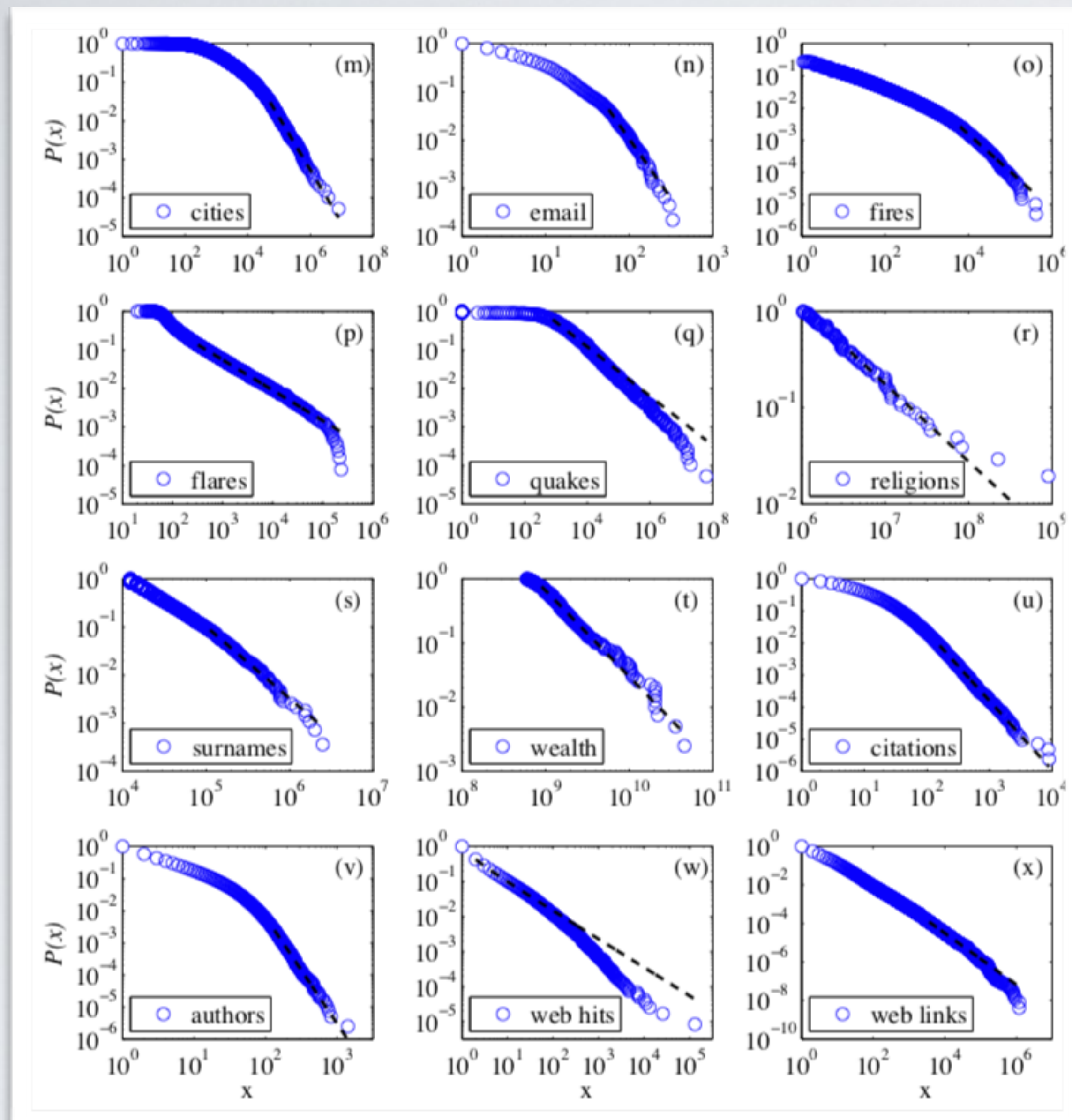
DEGREE DISTRIBUTION

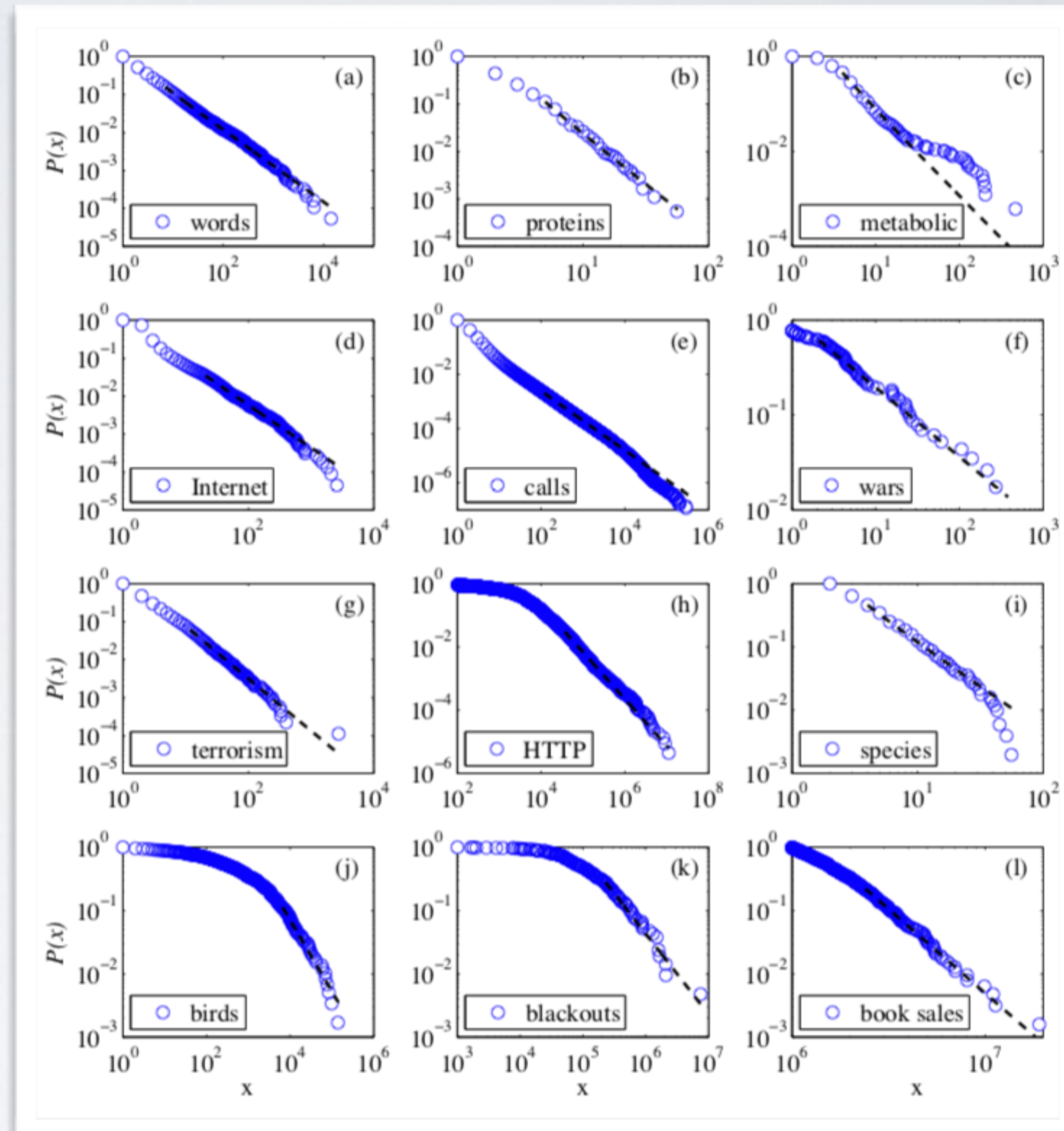
- In a fully random graph (Erdos-Renyi), degree distribution is a normal distribution centered on the average degree

- In real graphs, in general, it is not the case:
  - A high majority of small degree nodes
  - A small minority of nodes with very high degree (Hubs)

- Often modeled by a **power law**

DEGREE DISTRIBUTION





Power laws in empirical data (degrees and other things)

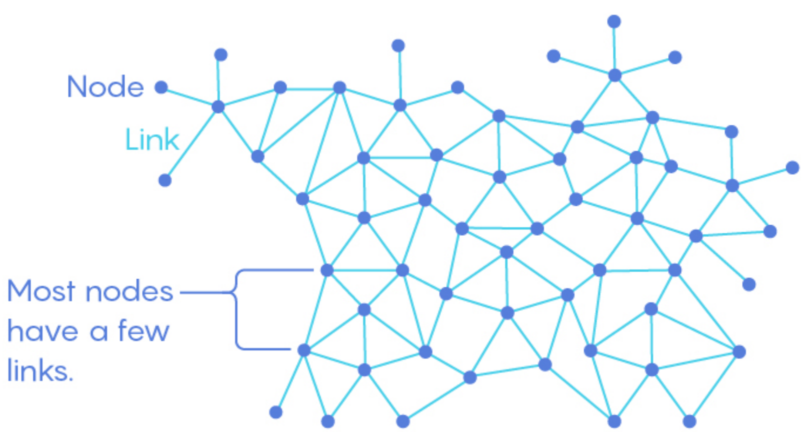
POWER-LAW DISTRIBUTIONS IN EMPIRICAL DATA [Clauset 2009]

To Be or Not to Be Scale-Free

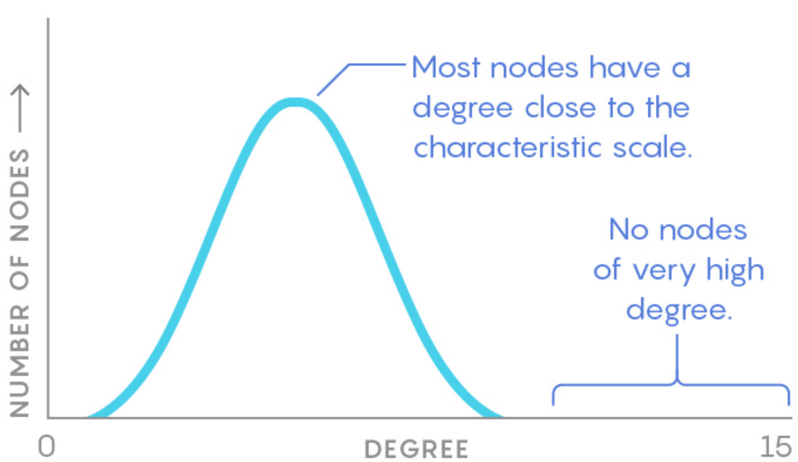
Scientists study complex networks by looking at the distribution of the number of links (or "degree") of each node. Some experts see so-called scale-free networks everywhere. But a new study suggests greater diversity in real-world networks.

Random Network

Randomly connected networks have nodes with similar degrees. There are no (or virtually no) "hubs" — nodes with many times the average number of links.

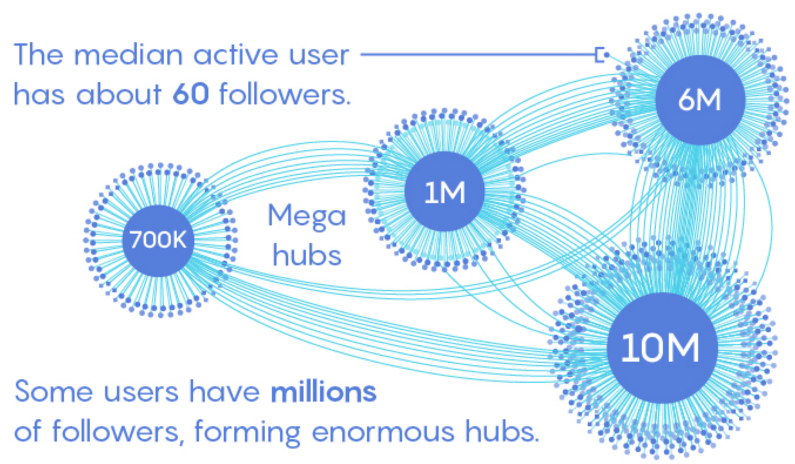


The distribution of degrees is shaped roughly like a bell curve that peaks at the network's "characteristic scale."

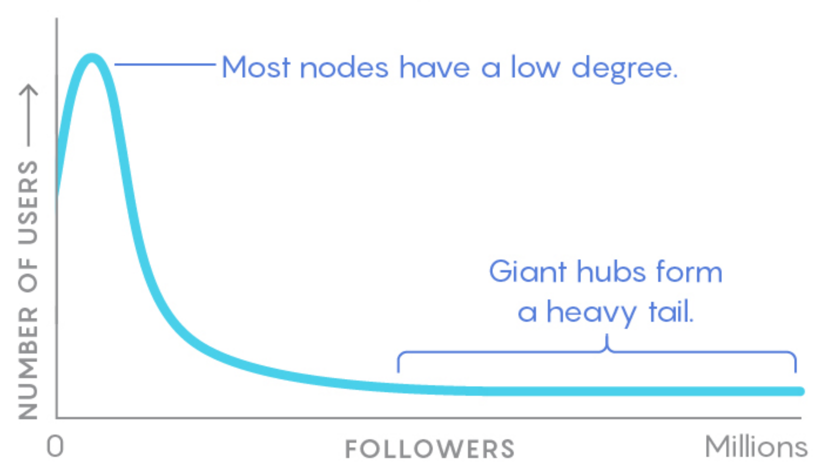


Twitter's Scale-Free Network

Most real-world networks of interest are not random. Some nonrandom networks have massive hubs with vastly higher degrees than other nodes.

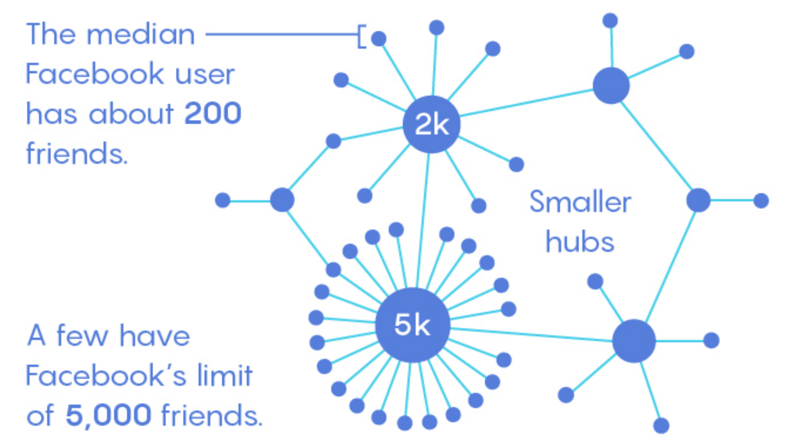


The degrees roughly follow a power law distribution that has a "heavy tail." The distribution has no characteristic scale, making it scale-free.

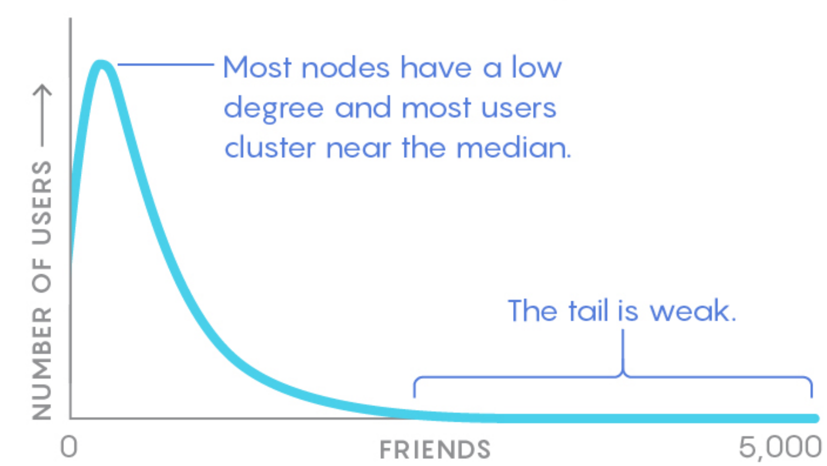


Facebook's In-Between Network

Researchers have found that most nonrandom networks are not strictly scale-free. Many have a weak heavy tail and a rough characteristic scale.



This network has fewer and smaller hubs than in a scale-free network. The distribution of nodes has a scale and does not follow a pure power law.



[Quanta magazine 2018]

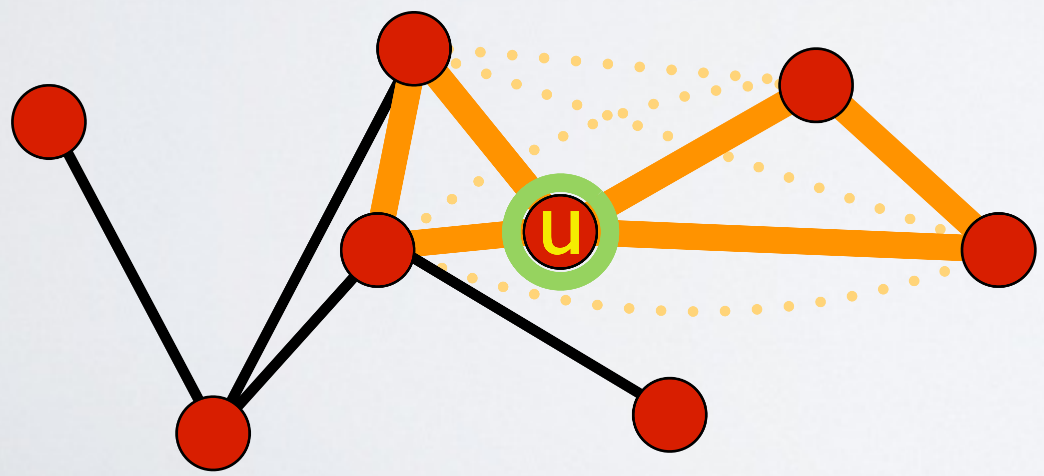
Node clustering coefficient

- Measure of interconnectivity
- What fraction of neighbours of a node are connected to each other?

Global clustering coefficient

$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triples of vertices}} = $$

$$= \frac{\text{number of closed triplets}}{\text{number of connected triples of vertices}}.$$



C = 9/18 = 1/2

$C_u = (2x2)/(4x3) = 1/3$

Local clustering coefficient

$$C_u = \frac{2e_u}{k_u(k_u - 1)}$$

- e_u - number of links between the neighbours of node u
- $(k_u(k_u\text{-}1))/2$ - maximum number of triangles

Average local clustering coefficient

$$C = \frac{1}{N} \sum_u C_u$$

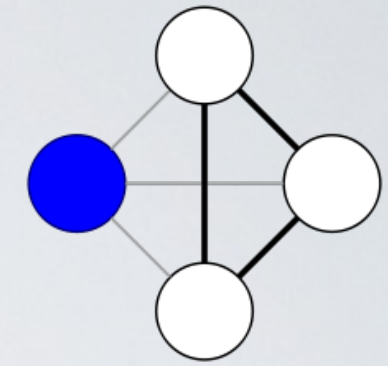
Definition: Watts and Strogatz 2002

CLUSTERING COEFFICIENT

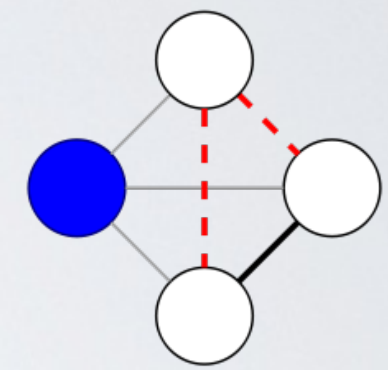
The higher the value,
the more **locally dense** is the network.

“Friends of my friends are my friends”

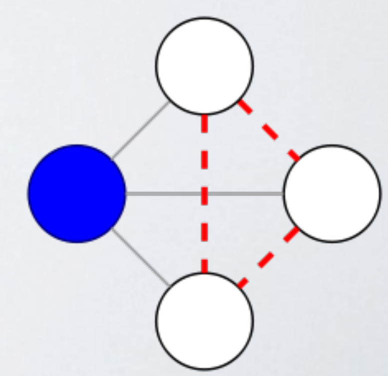
Higher in real networks than random



c = 1



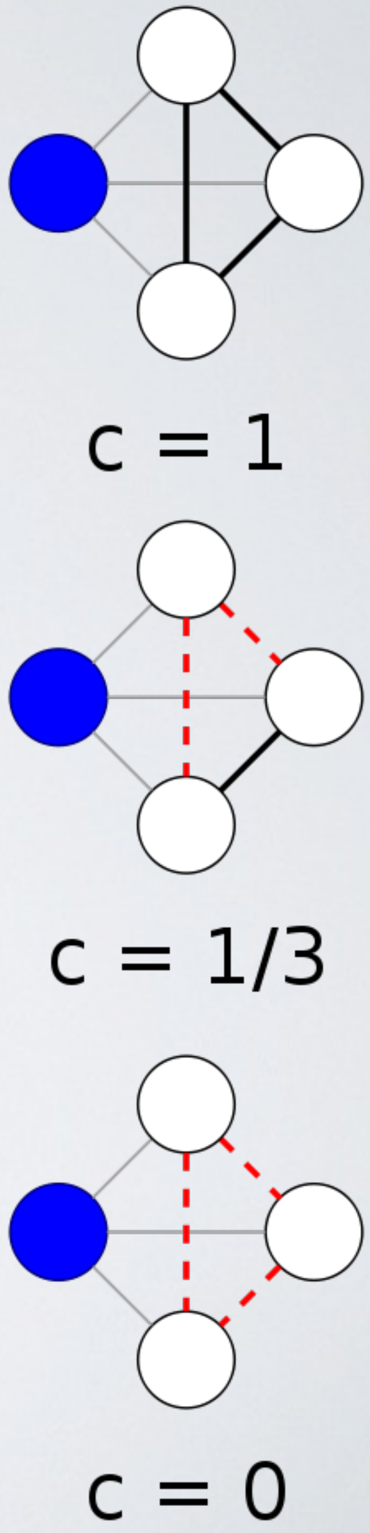
c = 1/3



c = 0

CLUSTERING COEFFICIENT

- Global CC:
  - Random (ER): =density: very small for large graphs
  - Facebook ego-networks: 0.6
  - Twitter lists: 0.56
  - California Road networks: 0.04

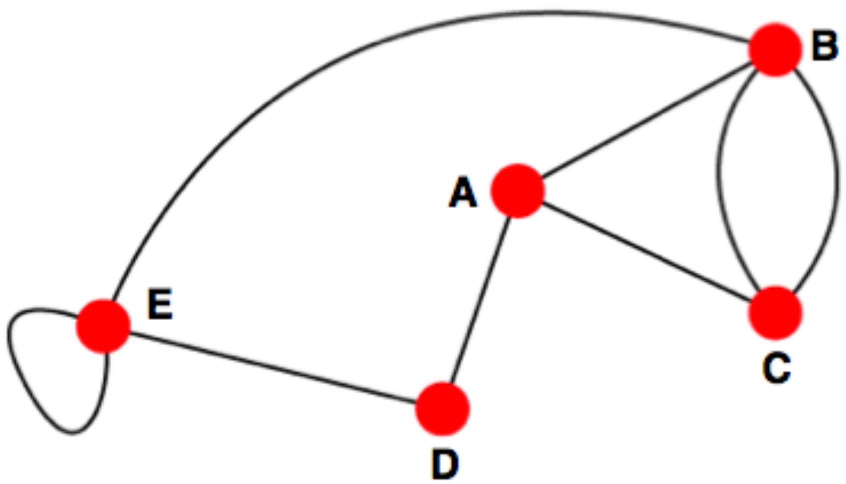


Path length

A ***path is*** a sequence of nodes in which each node is adjacent to the next one

$P_{i0,in}$ of length ***n*** between nodes i_0 and i_n is an ordered collection of ***n+1*** nodes and ***n*** links

$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$

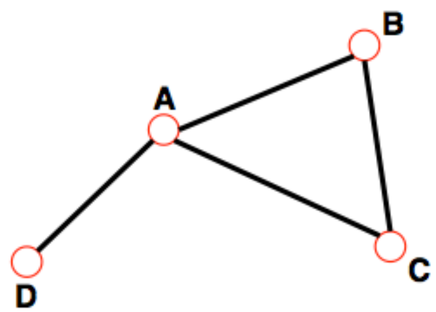


•A path can intersect itself and pass through the same link repeatedly. Each time a link is crossed, it is counted separately

•A legitimate path on the graph on the right:
ABCBCADEEBA

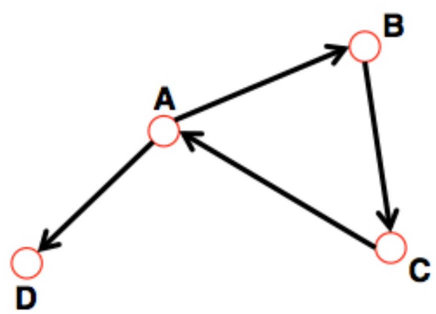
• In a directed network, the path can follow only the direction of an arrow.

Path length



The ***distance (shortest path, geodesic path)*** between two nodes is defined as the number of edges along the shortest path connecting them.

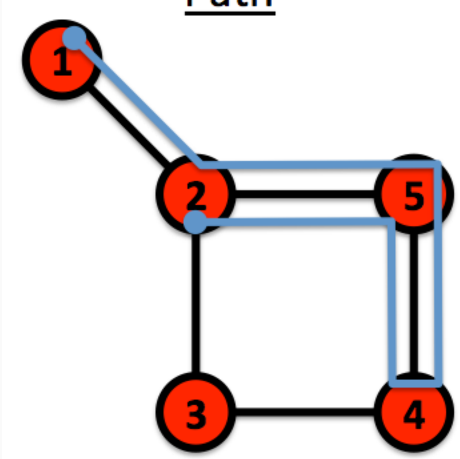
*If the two nodes are disconnected, the distance is infinity.



In directed graphs each path needs to follow the direction of the arrows.

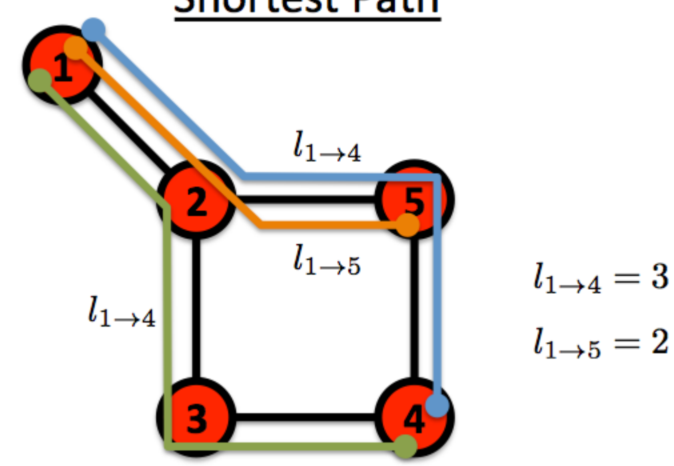
Thus in a digraph the distance from node A to B (on an AB path) is generally different from the distance from node B to A (on a BCA path).

Path



A sequence of nodes such that each node is connected to the next node along the path by a link.

Shortest Path



The path with the shortest length between two nodes (distance).

Path length

- d_{max} diameter- the maximum distance between any pairs of nodes

- $\langle d \rangle$ average path length - for directed graphs

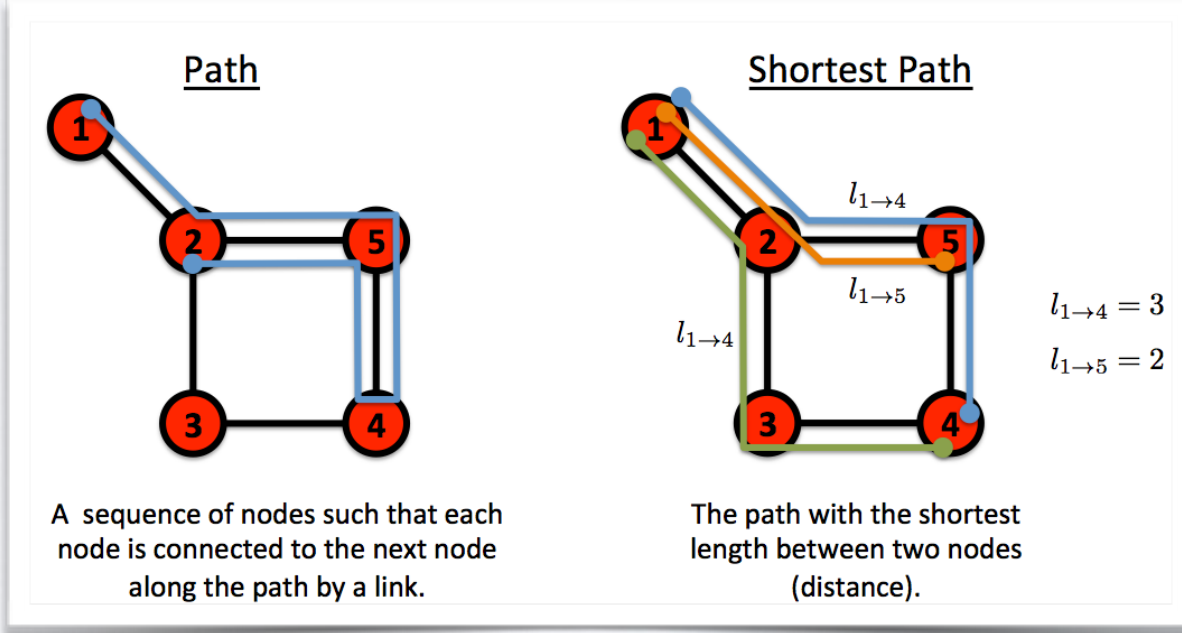
$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}$$

- where d_{ij} is the shortest distance between nodes i and j
- multiplicative is *(2 x max number of links)*
- distance between unconnected nodes is 0

- $\langle d \rangle$ average path length - for un-directed graphs

$$\langle d \rangle = \frac{2}{N(N-1)} \sum_{i < j} d_{ij}$$

- since $d_{ij} = d_{ji}$
- multiplicative is *(max number of links)*



AVERAGE PATH LENGTH

- The famous 6 degrees of separation (Milgram experiment)
  - In fact 6 hops
  - (More on that next slide)
- Not too sensible to noise
- Tells you if the network is “stretched” or “hairball” like

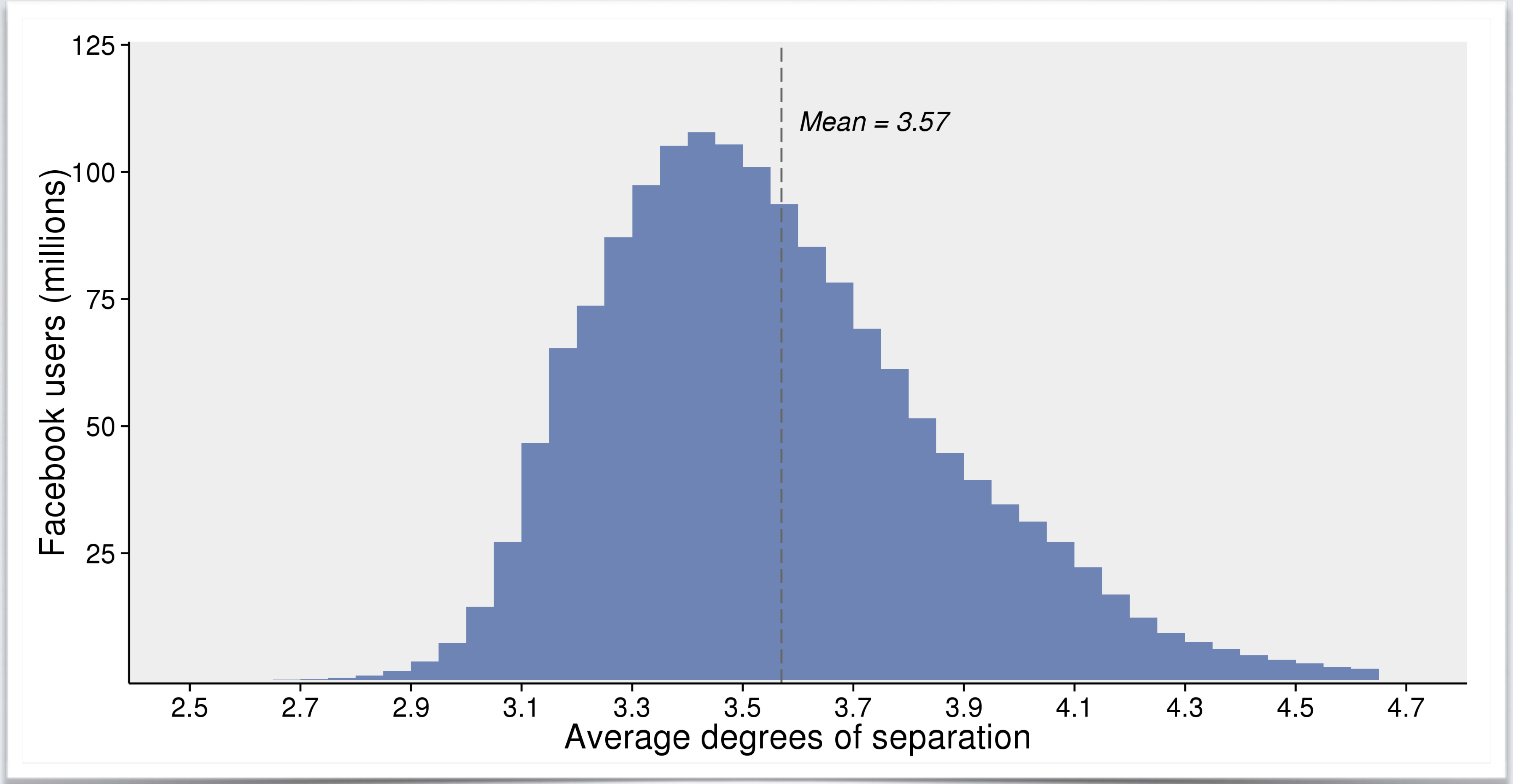
SIDE-STORY: MILGRAM EXPERIMENT

- Small world experiment (60's)
  - Give a (physical) mail to random people
  - Ask them to send to someone they don't know
    - They know his city, job
  - They send to their most relevant contact
- Results: In average, 6 hops to arrive

SIDE-STORY: MILGRAM EXPERIMENT

- Many criticism on the experiment itself:
  - Some mails did not arrive
  - Small sample
  - …
- Checked on "real" complete graphs (giant component):
  - MSN messenger
  - Facebook
  - The world wide web
  - …

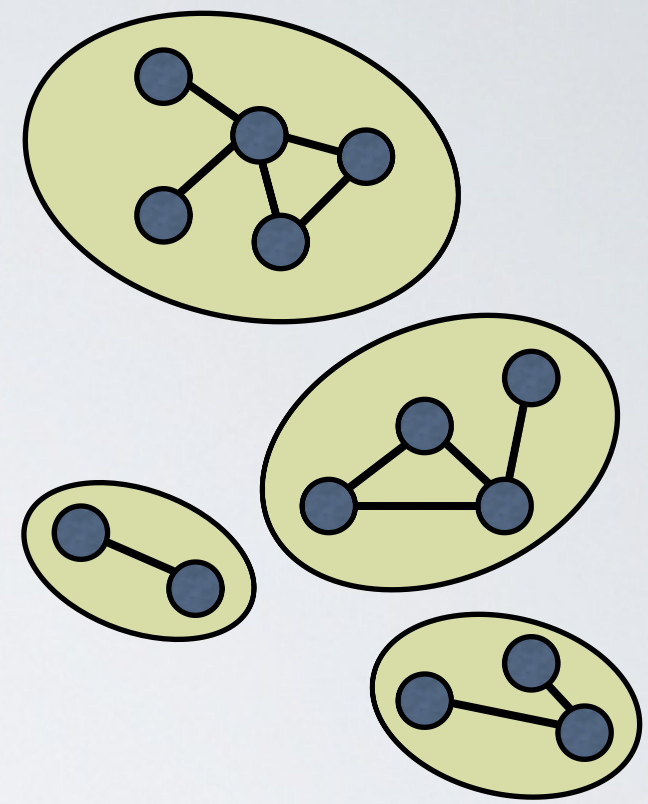
SIDE-STORY: MILGRAM EXPERIMENT



Facebook

Connectivity and components

- A connected component is a subset of vertices with at least one path connecting each of them
- A network may consist of a single connected component (a connected network) or several of those
- Distances between nodes in disjoint components are not defined (infinite)
- Bridge: if we remove it, the graph becomes disconnected.
- The adjacency matrix of a network with several components can be written in a block-diagonal form, so that nonzero elements are confined to squares, with all other elements being zero



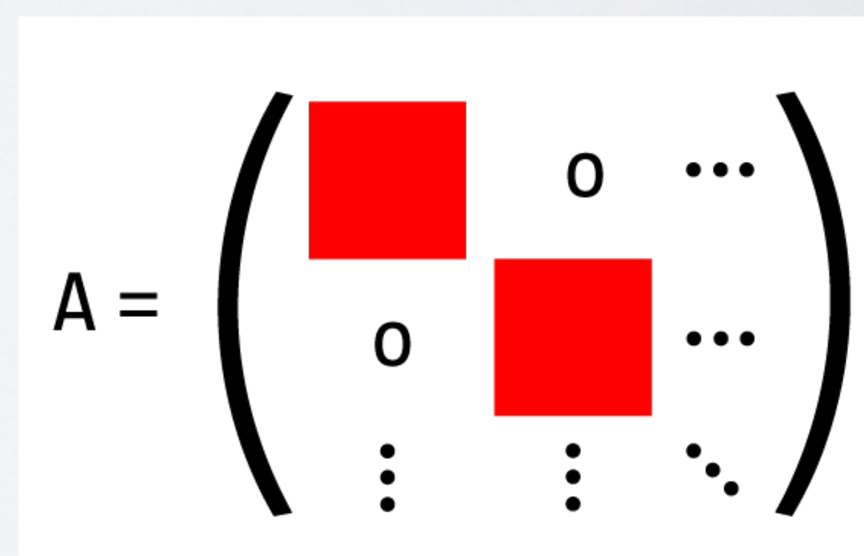


Figure after Newman, 2010

Connectivity and components - directed networks

- Strongly connected component (SCC): has a path from each node to every other node in the component
- Weakly connected component (WCC): it is connected if we disregard the directions
- In-component: nodes that can reach the SCC
- Out-component: nodes that can be reached from SCC

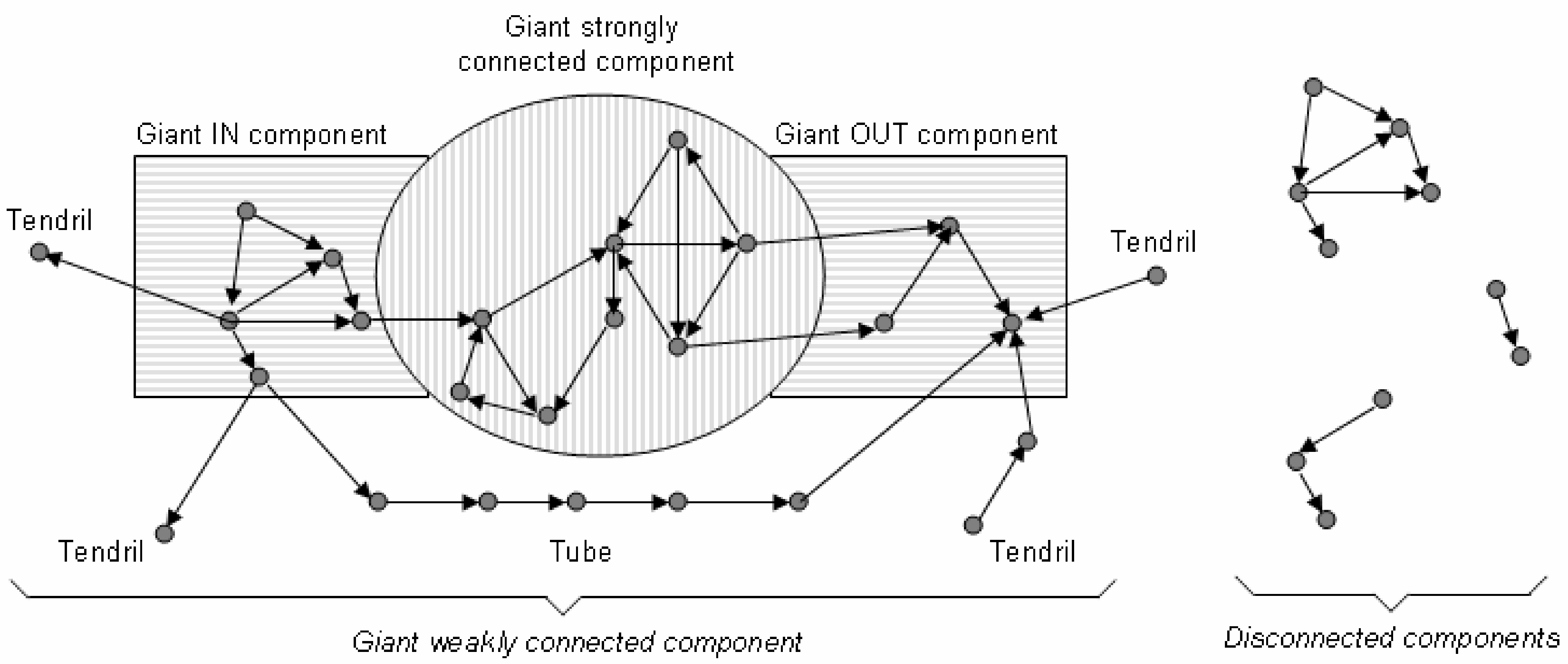


Figure from Broder et. al. (2000)

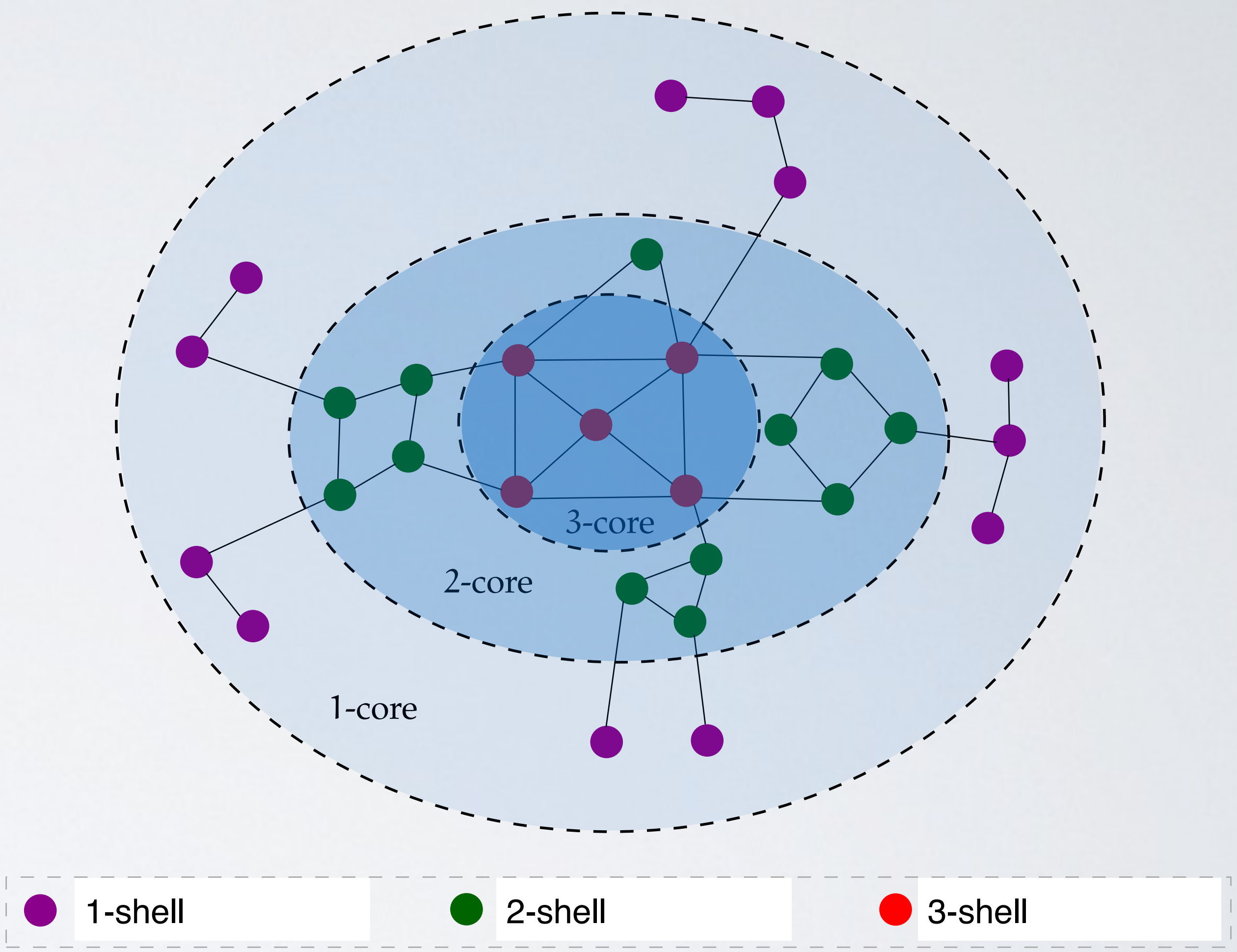
k-core decomposition

Goal: To identify dense cores of high degree nodes in networks

Given graph $G = (V, E)$

Definition: A subgraph $H = (C, E|C)$ induced by the set $C \subseteq V$ is a **k-core or a core of order k** iff $\forall v \in C : degree(H(v)) \geq k$, and H is the maximum subgraph with this property.

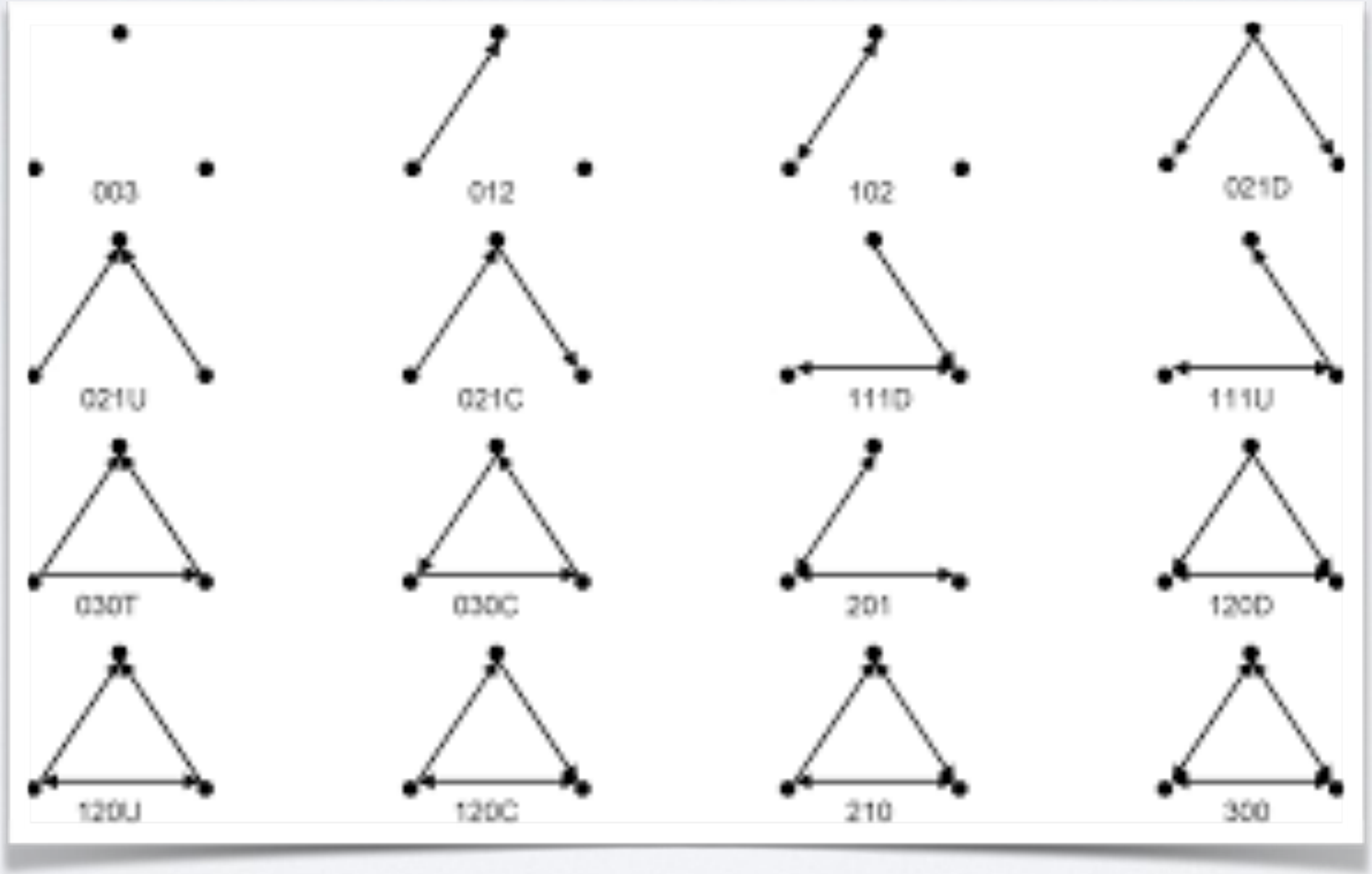
- A k-core of G can be obtained by recursively removing all the vertices of degree less than k, until all vertices in the remaining graph have at least degree k.



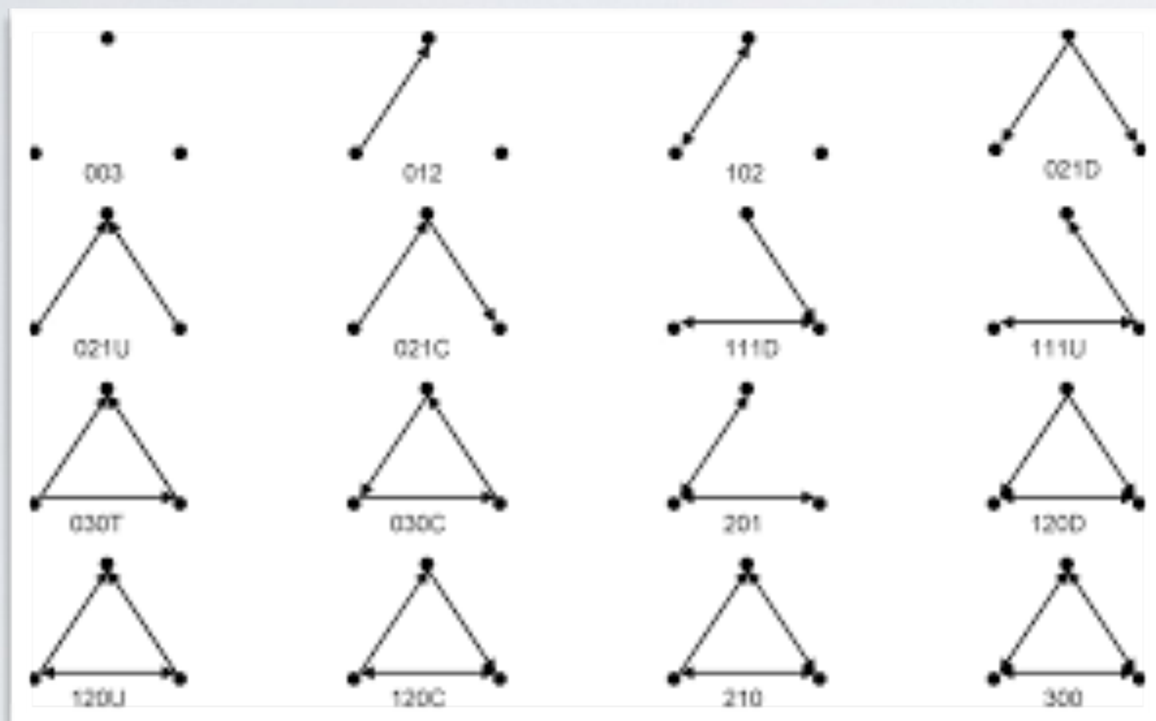
Definition: A vertex i has **coreness** $\boldsymbol{c}$ if it belongs to the c*-core* but not to $(c + 1)$*-core*.

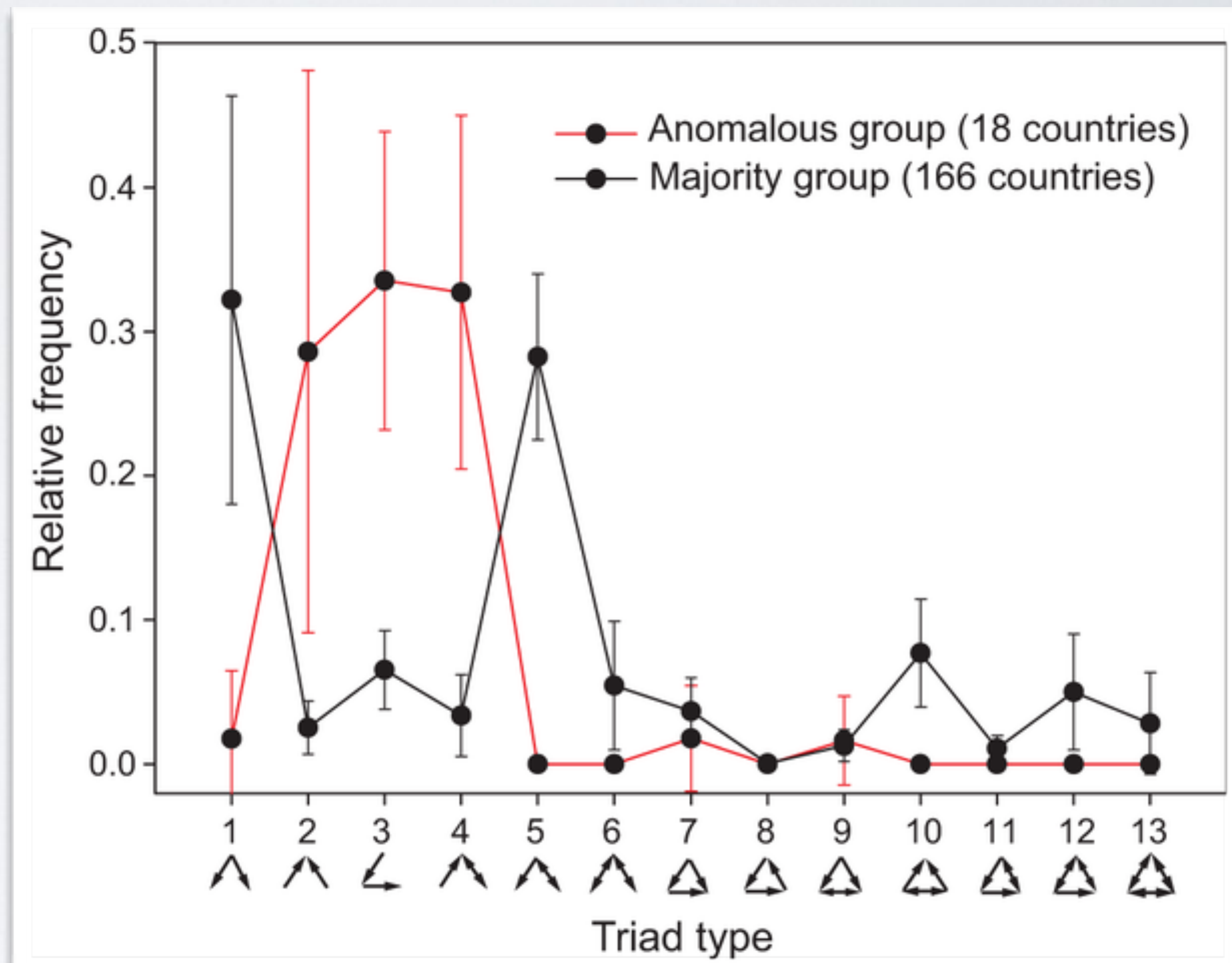
Definition: A $\boldsymbol{c}$**-shell** is composed by all the vertices whose coreness is c. The k-core is thus the union of all shells with $c \geq k$.

TRIADS COUNTING

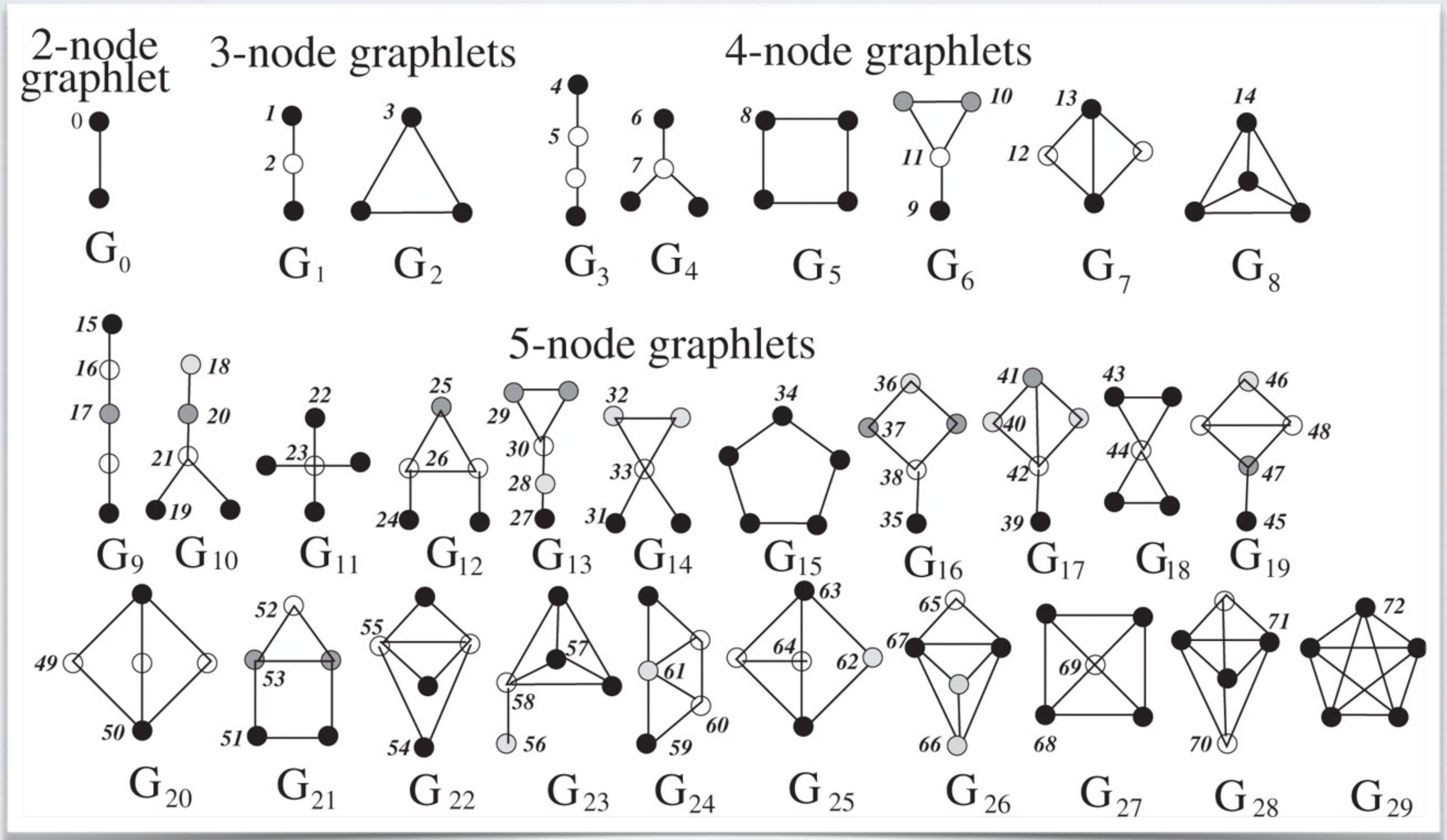


TRIADS COUNTING

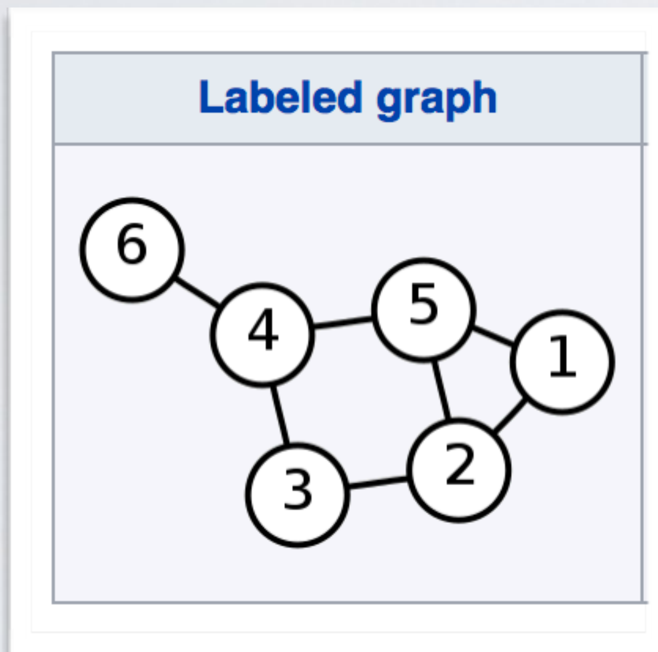




GRAPHLETS


2-node graphlet
3-node graphlets
4-node graphlets
5-node graphlets
G_0 G_1 G_2 G_3 G_4 G_5 G_6 G_7 G_8
G_9 G_{10} G_{11} G_{12} G_{13} G_{14} G_{15} G_{16} G_{17} G_{18} G_{19}
G_{20} G_{21} G_{22} G_{23} G_{24} G_{25} G_{26} G_{27} G_{28} G_{29}

MATRIX PROPERTIES



Adjacency matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

- What is a Matrix?
  - <u>Not</u> a 2D data table
  - It describes a *linear transformation*, or *linear function*
  - Said differently, it represents a *set of equations*

MATRIX PROPERTIES

$$
\begin{array}{c|c}
 & \begin{matrix} x1 & x2 & x3 & x4 & x5 & x6 \end{matrix} \\
\begin{matrix} x1' \\ x2' \\ x3' \\ x4' \\ x5' \\ x6' \end{matrix} &
\begin{pmatrix}
0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0
\end{pmatrix}
\end{array}
$$

$$x1' = 0x_1 + 1x_2 + 0x_3 + 0x_4 + 1x_5 + 0x_6$$
$$x2' = x_1 + x_3 + x_5$$
$$x3' = x_2 + x_4$$

$$\ldots$$

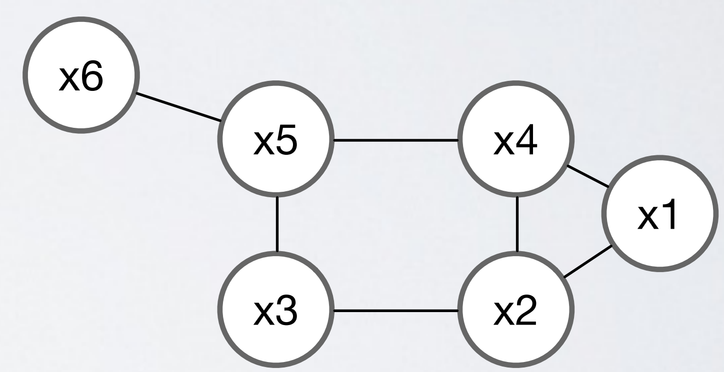
MATRIX PROPERTIES

$$A= \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

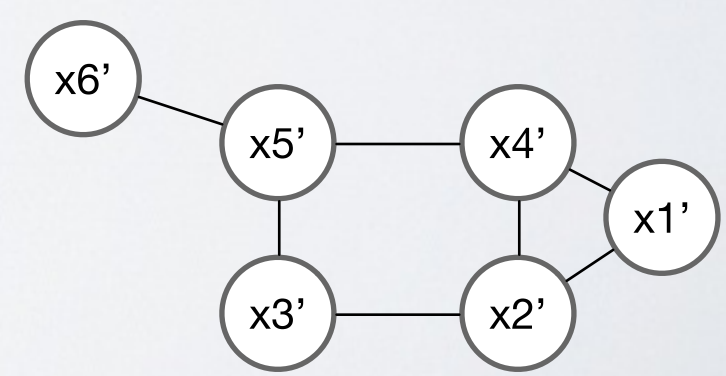
$$x= \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_5 \\ x_5 \\ x_6 \end{pmatrix}$$

$$Ax= \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_5 \\ x_5 \end{pmatrix} \begin{pmatrix} x_2 + x_4 \\ x_1 + x_3 + x_5 \\ x_2 + x_4 \\ x_3 + x_5 + x_6 \\ x_1 + x_2 + x_4 \\ x_4 \end{pmatrix}$$

A, x =



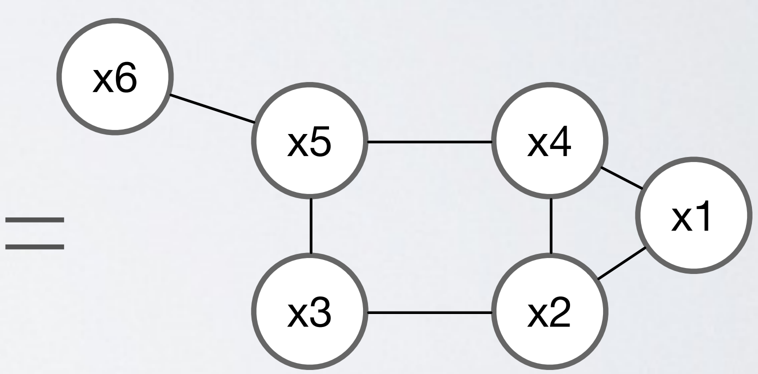
A, Ax=



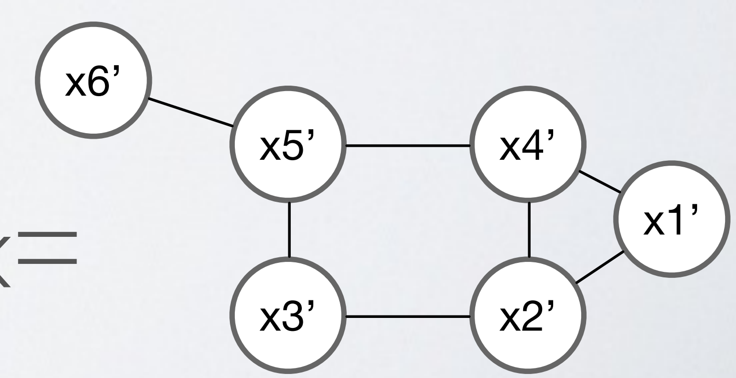
MATRIX PROPERTIES

- Question: What is the result of Ax if
  - x1=x2=x3=x4=x5=x6=1 ?

A, x =



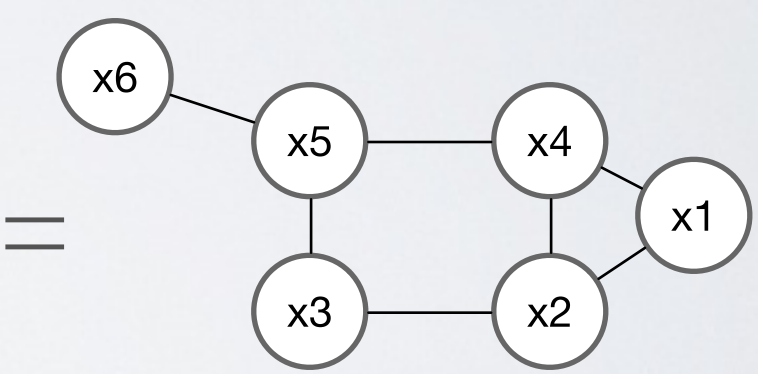
A, Ax=



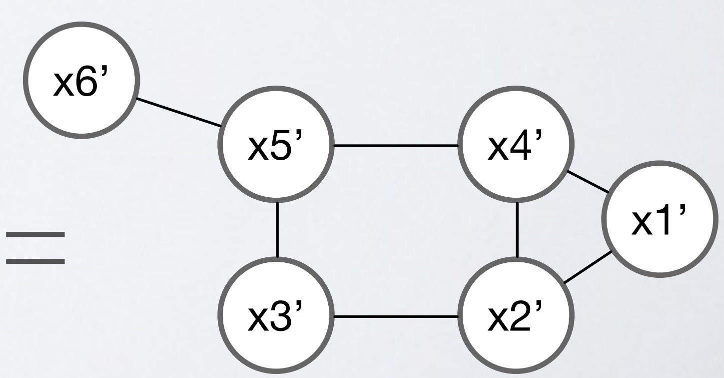
MATRIX PROPERTIES

- Question: What is the result of Ax if
  - x1=x2=x3=x4=x5=x6=1 ?
  - =>New values are degrees

A, x =



A, Ax=



MATRIX PROPERTIES

- What about A^2 ?
  - A encodes the *number of paths of lengths exactly* ***1*** *between pairs of nodes*
  - A^2 encodes the *number of paths of lengths exactly* ***2*** *between pairs of nodes*
  - A^3 encodes the *number of paths of lengths exactly* ***3*** *between pairs of nodes*
  - …

- Graph matrices operations can be interpreted as:
  - Diffusion phenomenons
  - Random walks

EXEMPLE OF GRAPH ANALYSIS

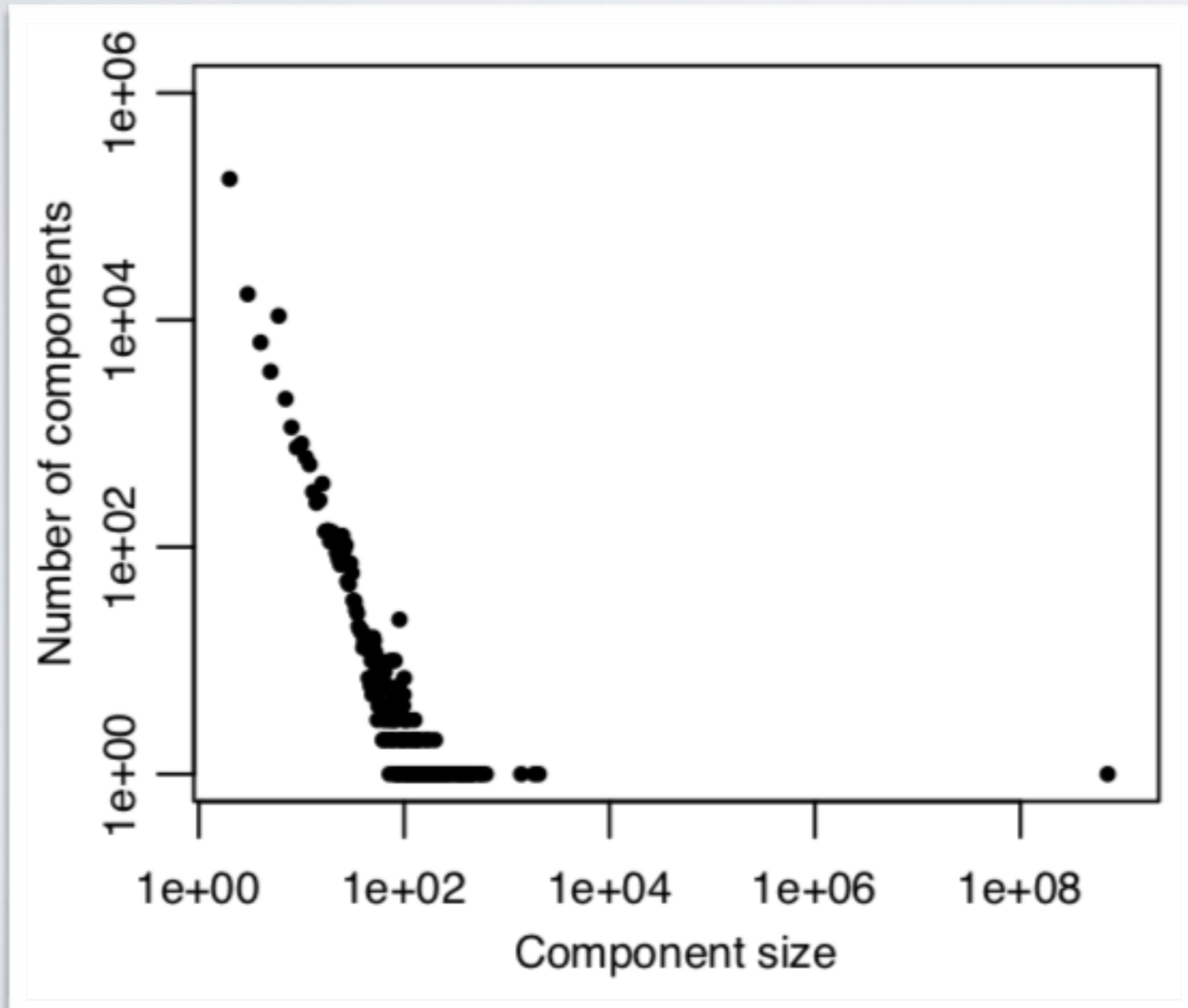
- Source: [The Anatomy of the Facebook Social Graph, Ugander et al. 2011]

- The Facebook friendship network in 2011

EXEMPLE OF GRAPH ANALYSIS

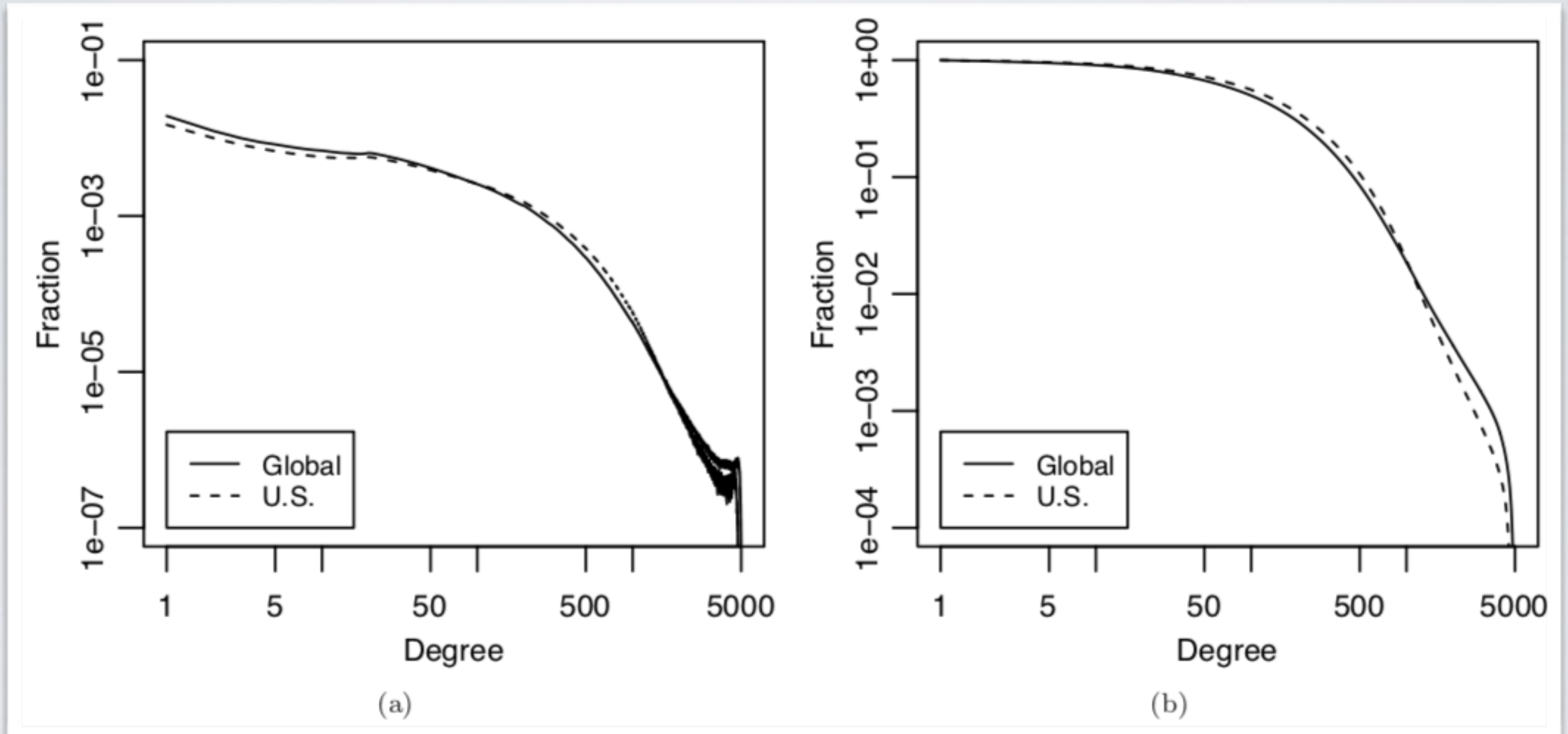
- 721M users (nodes) (active in the last 28 days)
- 68B edges
- Average degree: 190 (average # friends)
- Median degree: 99
- Connected component: 99.91%

EXEMPLE OF GRAPH ANALYSIS



Component size
Distribution

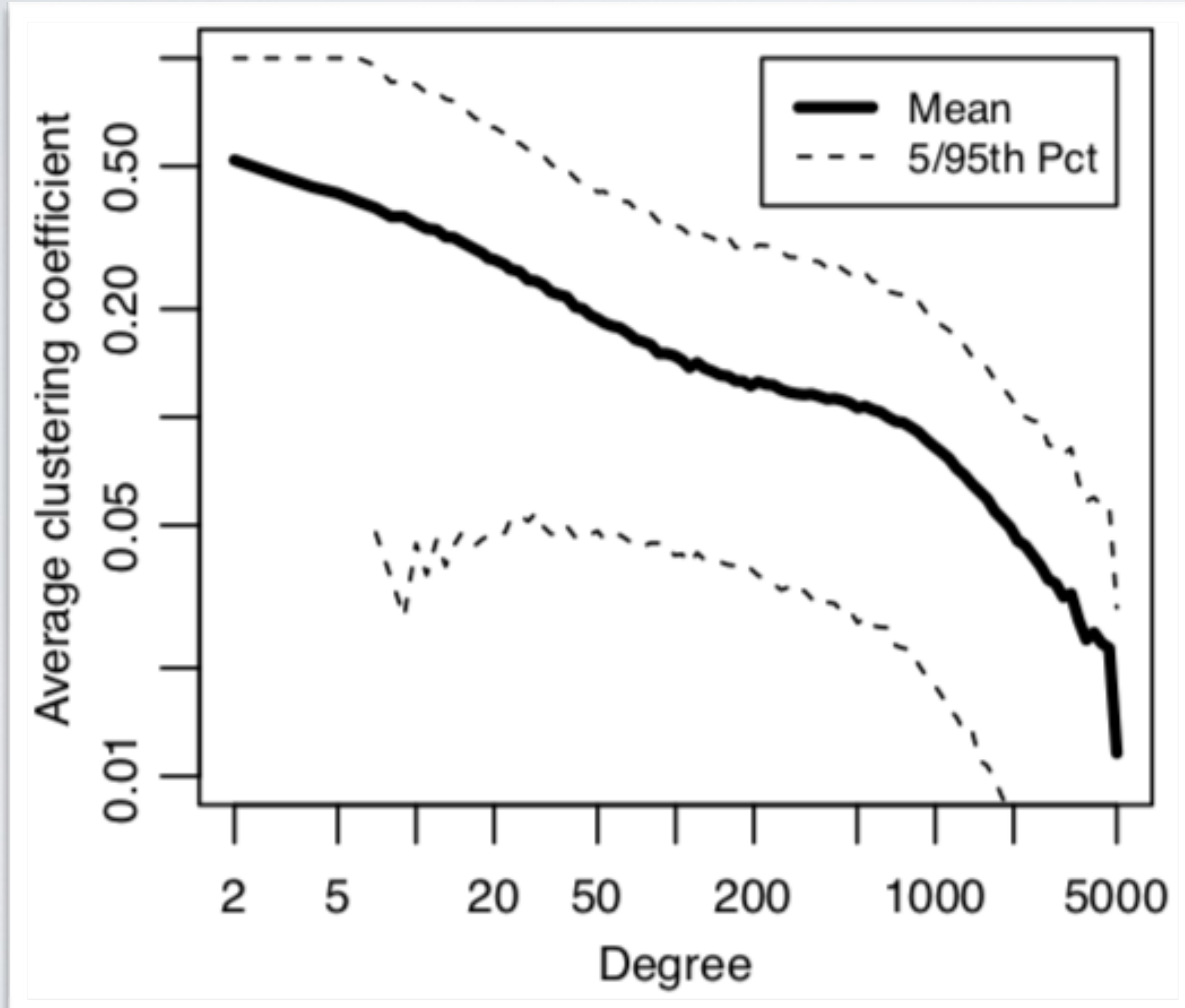
EXEMPLE OF GRAPH ANALYSIS



Cumulative

Degree distribution

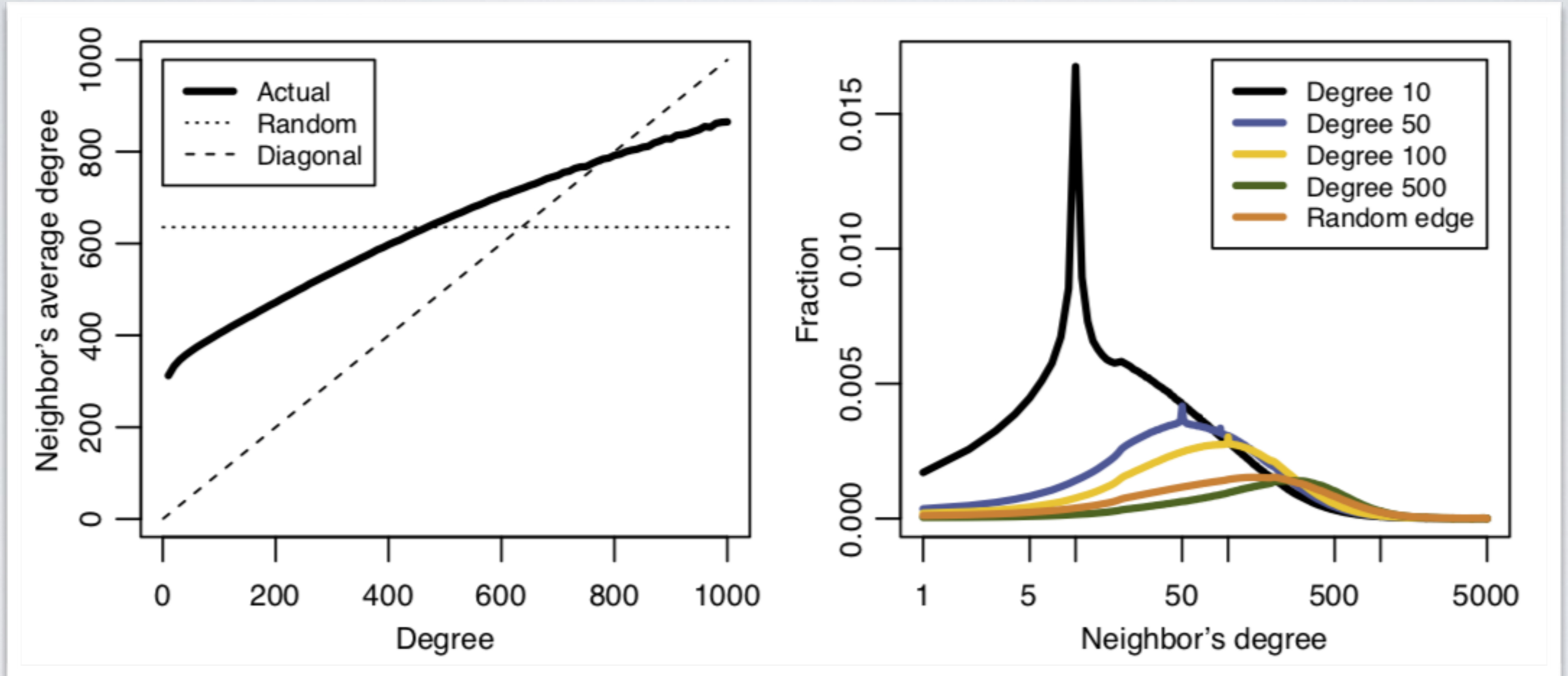
EXEMPLE OF GRAPH ANALYSIS



Clustering coefficient
By degree

Median user: 0.14:
14% of users with a common friend are friends

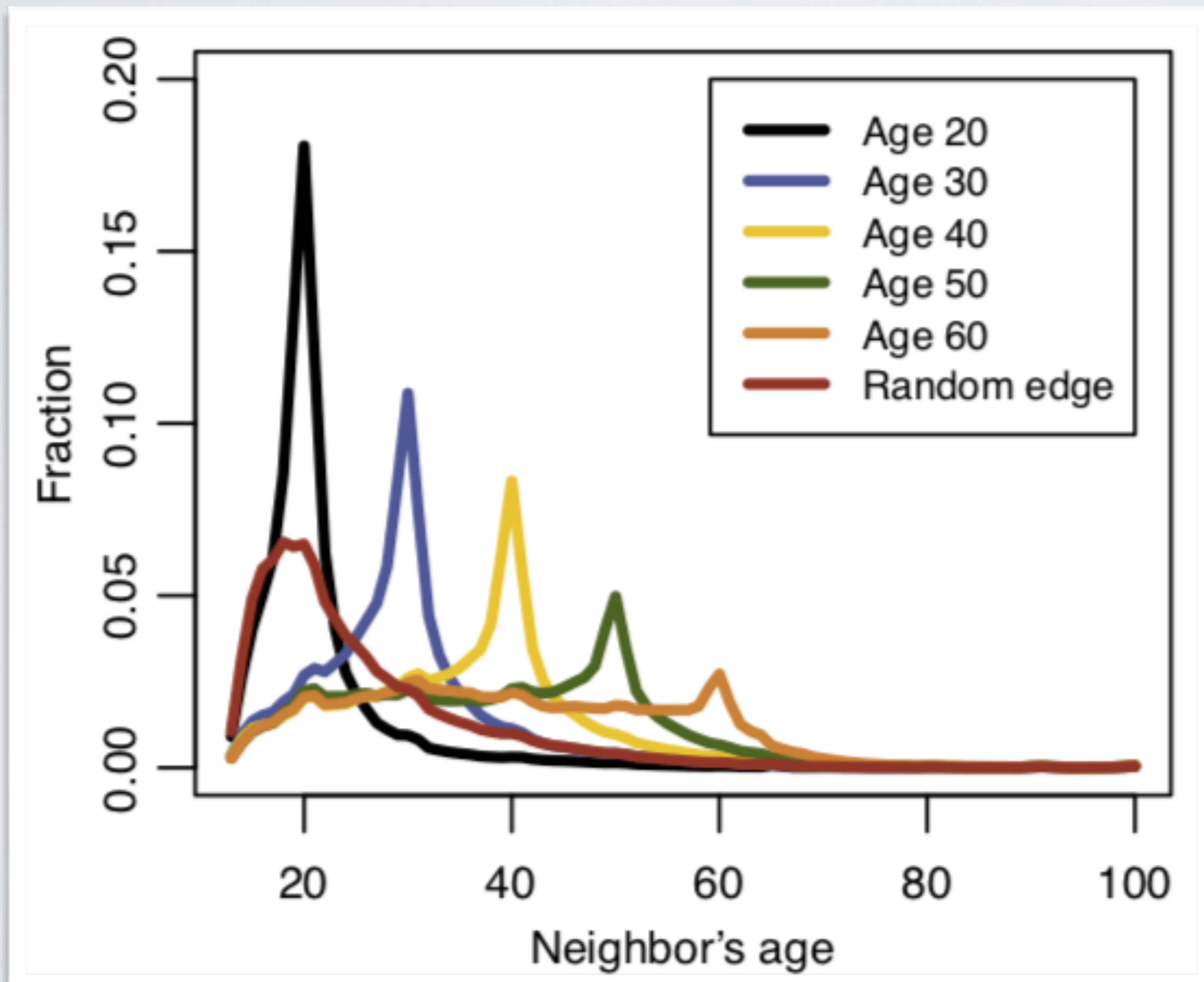
EXEMPLE OF GRAPH ANALYSIS



My friends have more
Friends than me!

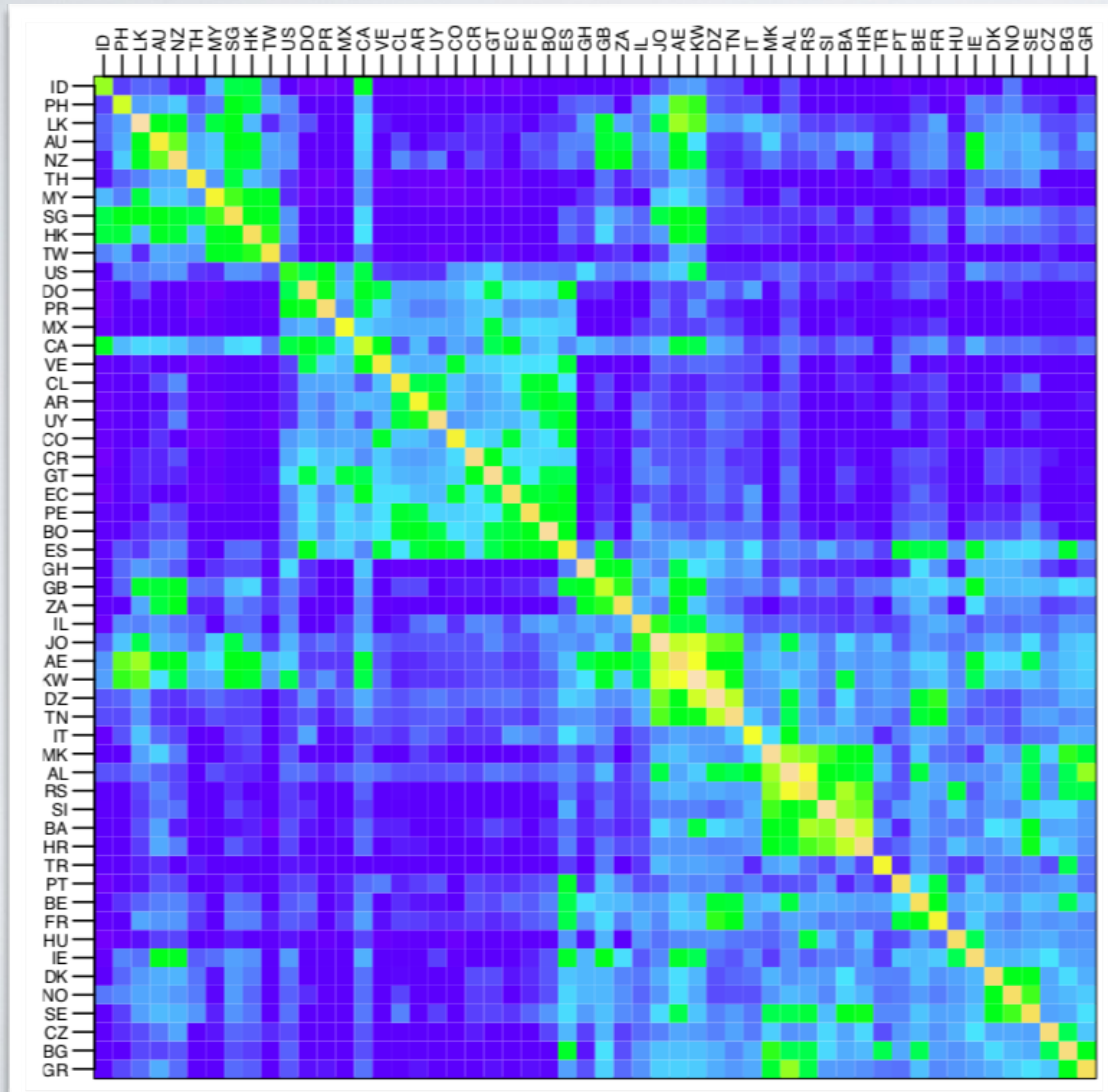
Many of my friends have the
Same # of friends than me!

EXEMPLE OF GRAPH ANALYSIS



Age homophily

EXEMPLE OF GRAPH ANALYSIS



Country similarity

84.2% percent of edges are within countries

(More in the community detection class)

NEXT CLASSES

- 1) Describe a network
- 2) Find and describe important nodes
- 3)Find and describe important group of nodes
  - And a few more things

PROJECT PRESENTATION

PROJECT OBJECTIVE

- We have in a database all transactions between addresses and all transactions between actors from the beginning of bitcoin to August 2016

- Choose and obtain a small subset of this network that you consider interesting
  - Around a particular transaction (illegal activity …)
  - About some actors
  - About a short period
  - …

PROJECT OBJECTIVE

- Apply tools your learned about during the class to better understand this network

- Write a report about what you learnt, and what you could learn with more time/data
  - If possible, a single Jupyter notebook with code and text
  - A separate report is also possible if relevant

PROJECT OBJECTIVE

- Recommendations:

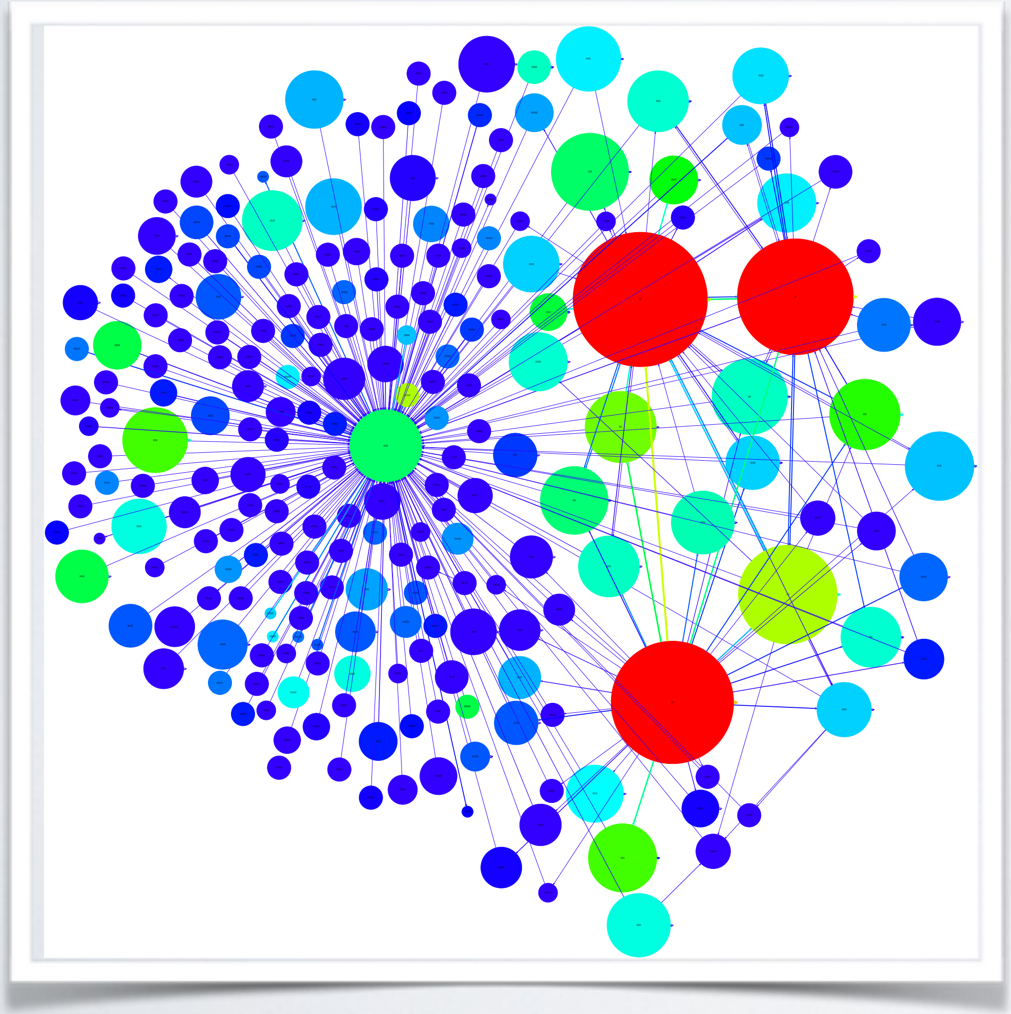
- I recommend to limit yourself to a few thousand nodes, and less than 10.000 edges

- The goal of the project is to interact!
  - Ask me if something is possible, how to do it… we are doing the project together.

SOME IDEAS

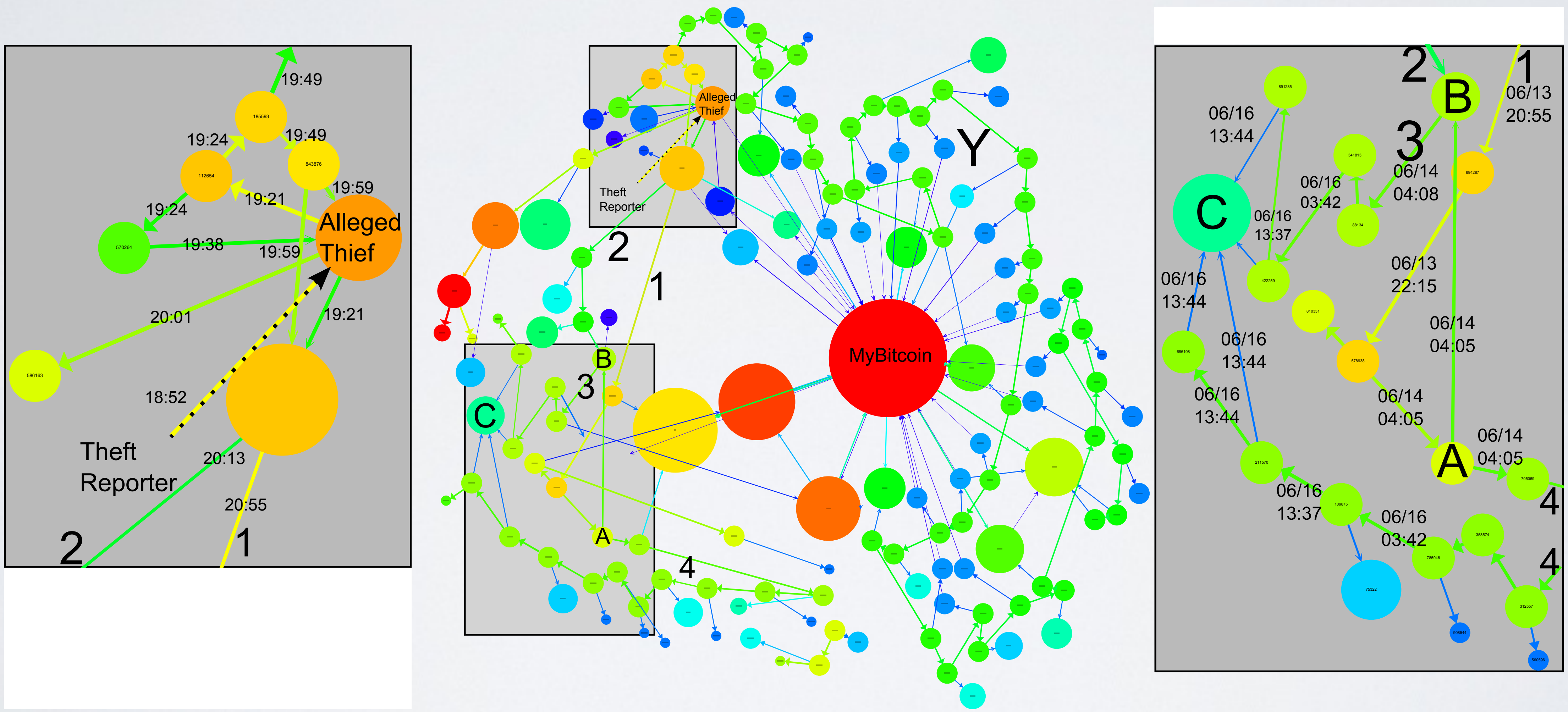
EGO-CENTERED NETWORK

- Wikileaks



Green in the center : wikileaks

A BITCOIN THEFT



MONEY LAUNDERING

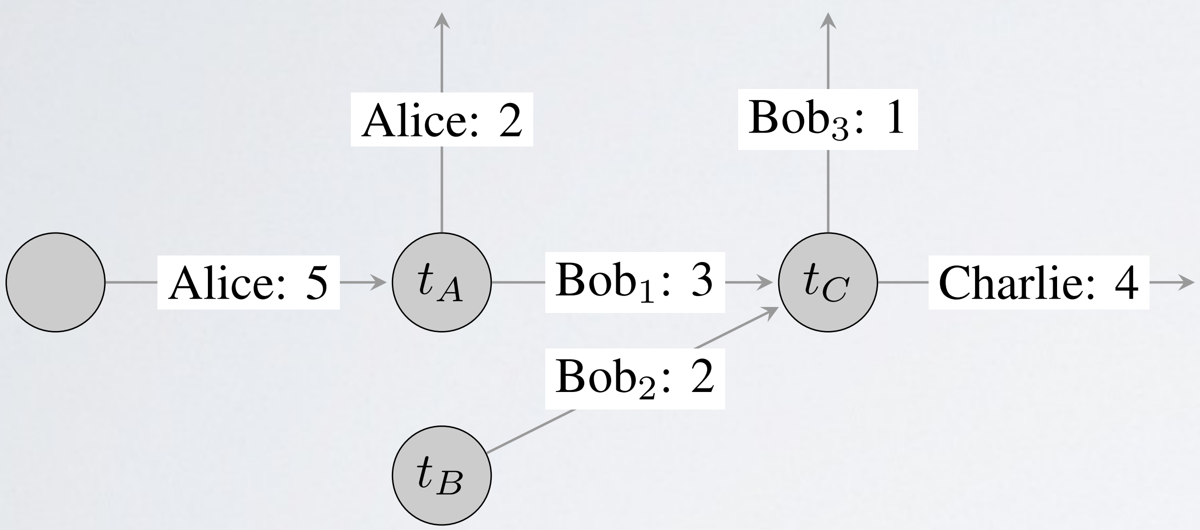


Figure 1. Example of a partial transaction graph

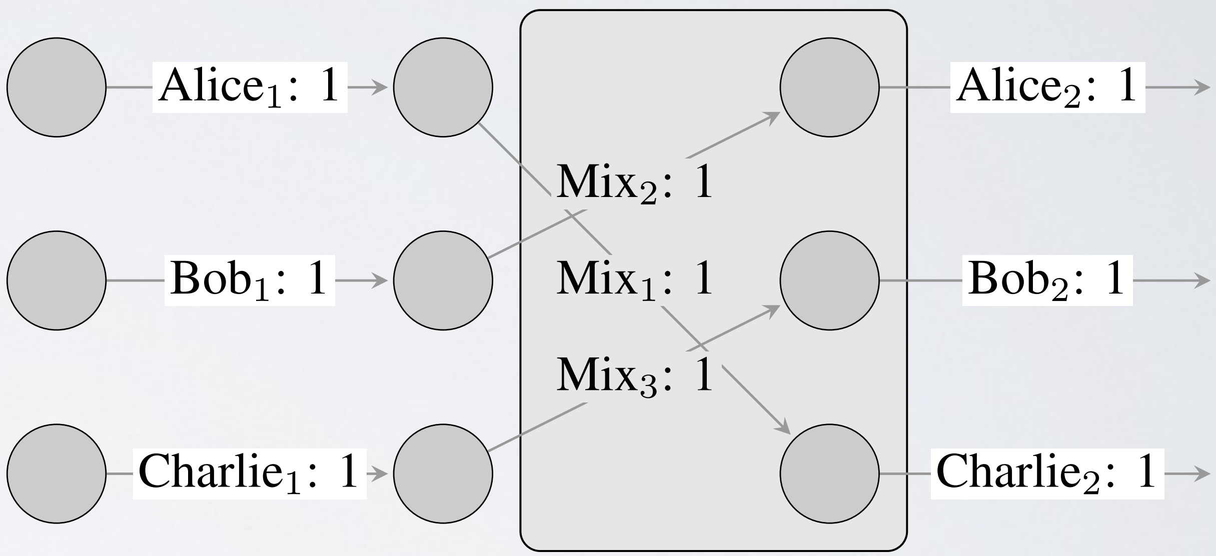
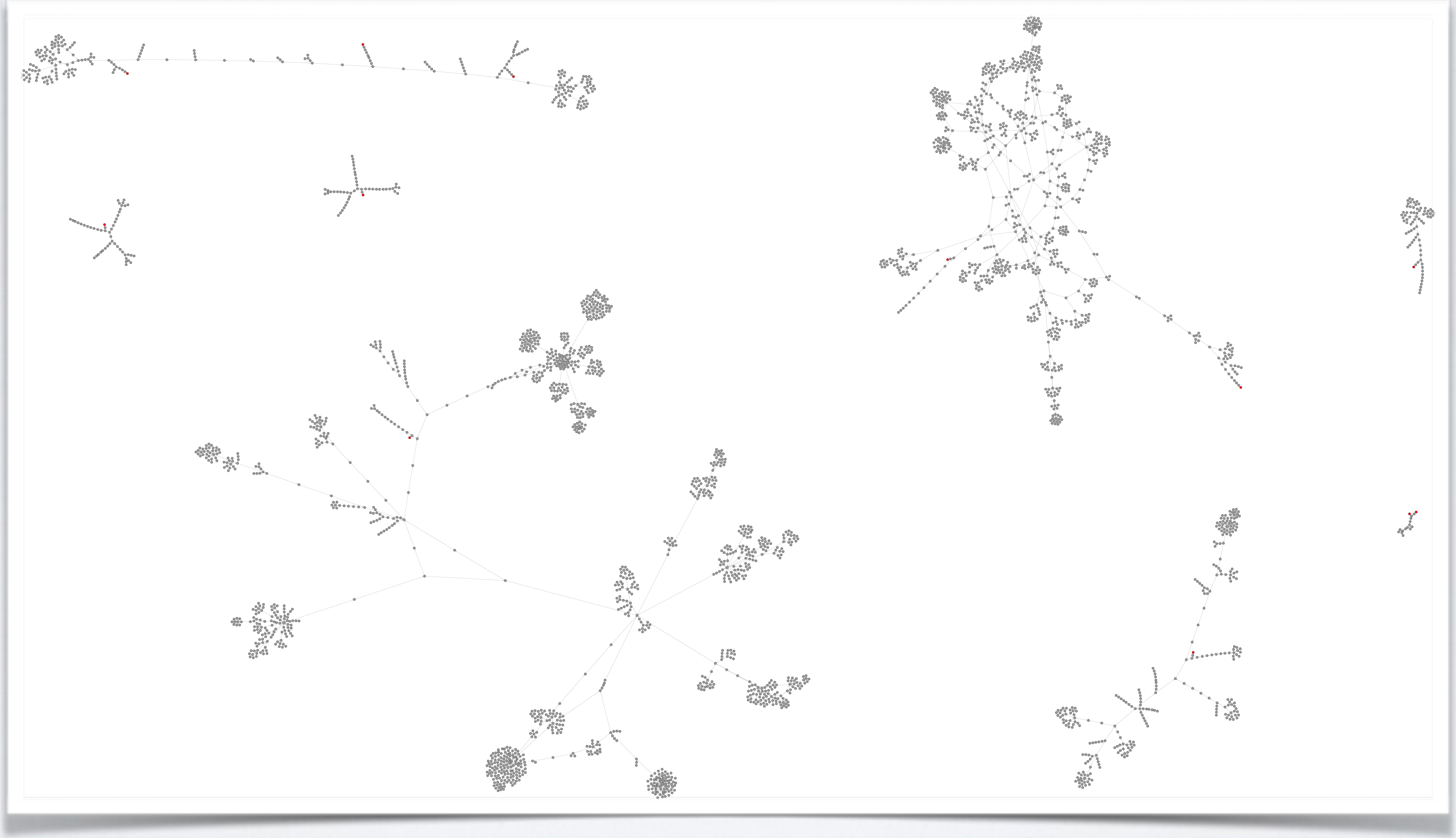


Figure 2. Block diagram of a hypothetical Bitcoin mixing service

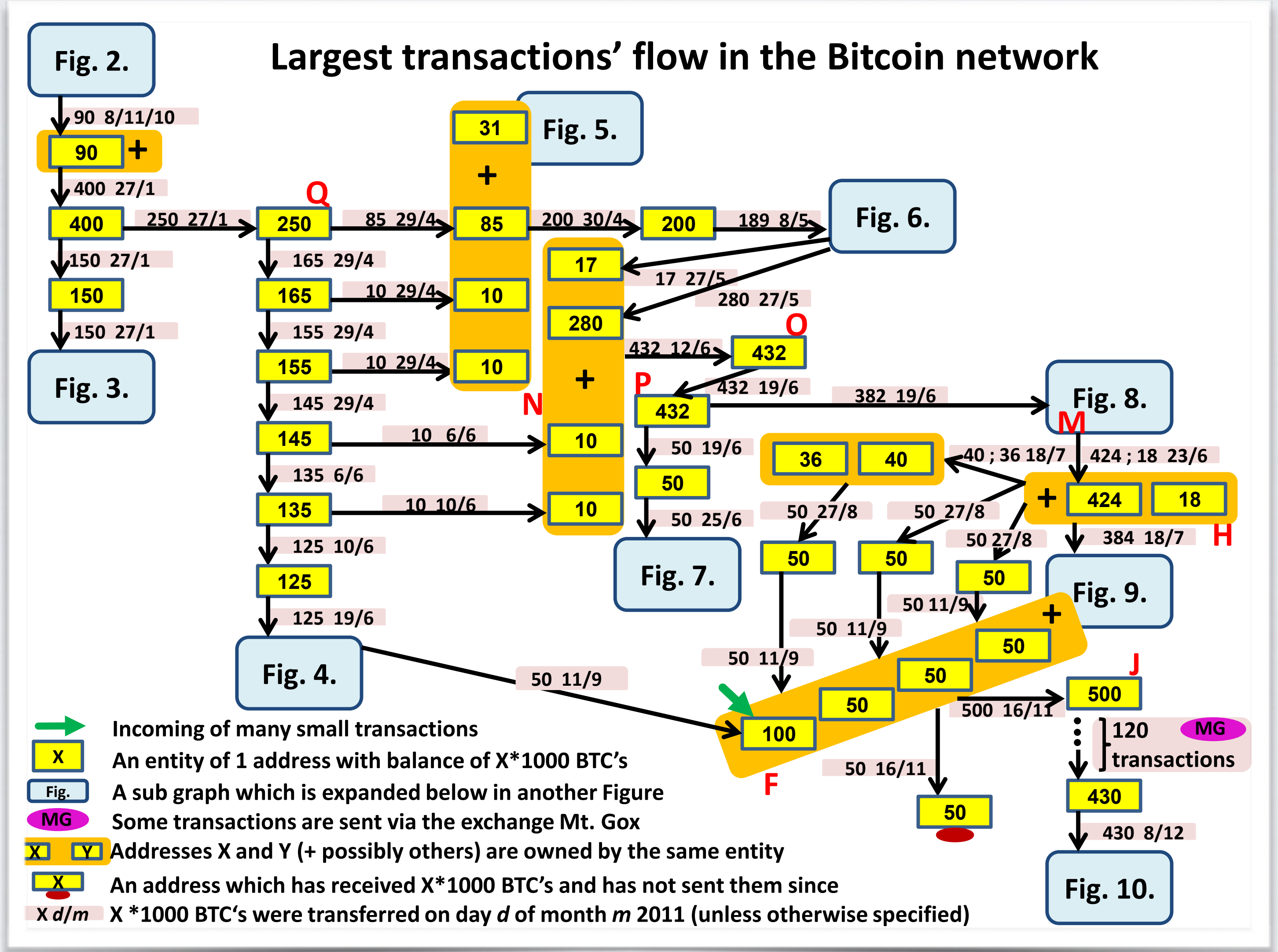
[Möser & Böhme]

MONEY LAUNDERING



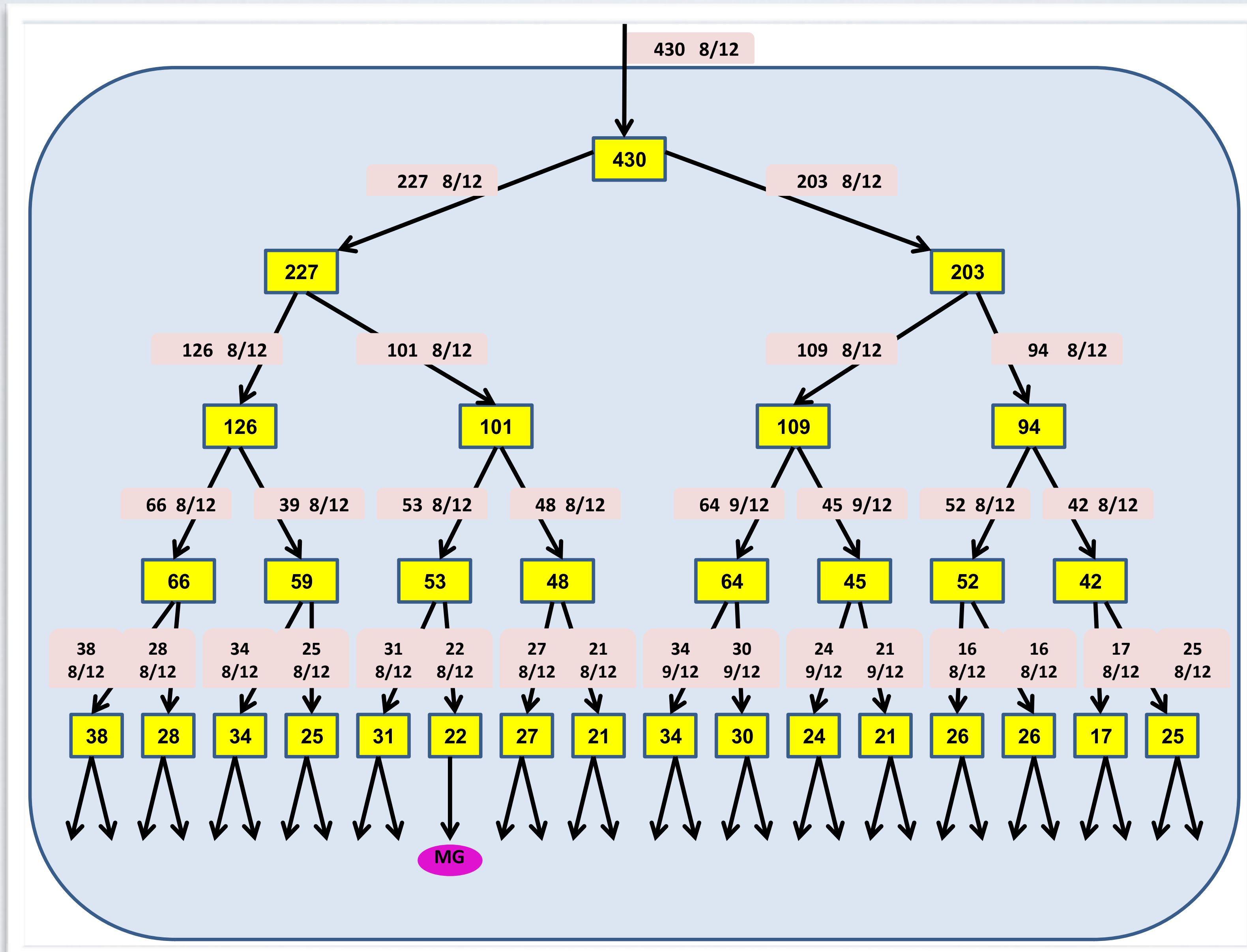
[Möser & Böhme]

EXCEPTIONAL TRANSACTIONS ANALYSIS



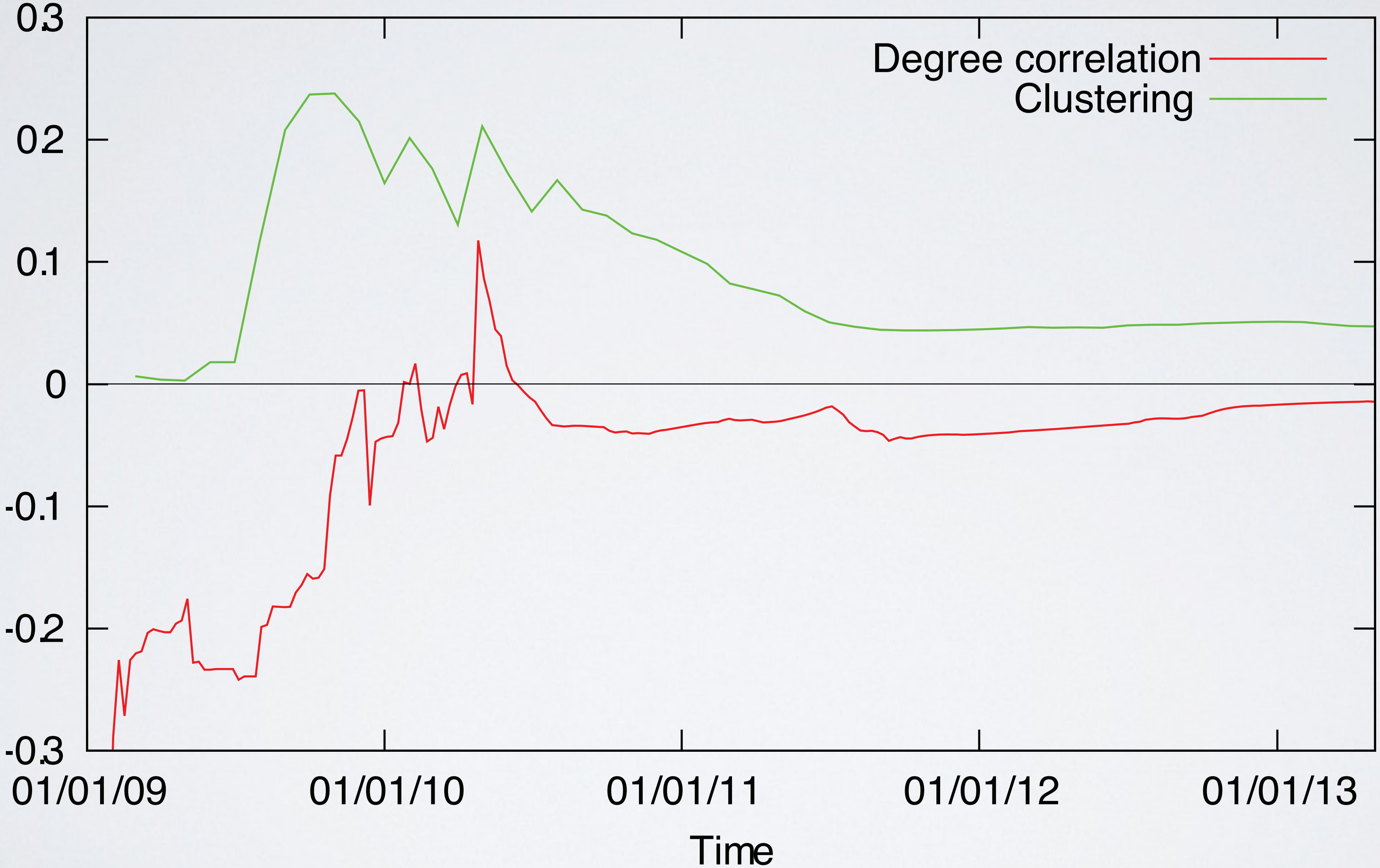
[Ron & Shamir]

EXCEPTIONAL TRANSACTIONS ANALYSIS



[Ron & Shamir]

ANALYSIS OF NETWORK PROPERTIES



[Kondor et al]

WHAT TO DO NOW

- http://cazabetremy.fr/Teaching/BitcoinNetwork.html
- Download the two provided networks. Choose one and load it with Gephi