RECSM Summer School: Social Media and Big Data Research

Pablo Barberá
London School of Economics
www.pablobarbera.com

Course website:

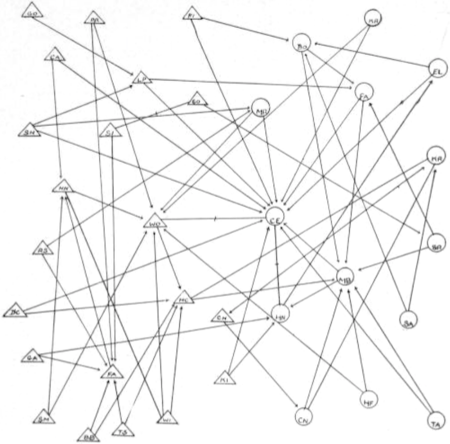
pablobarbera.com/social-media-upf

Discovery in Large-Scale Social Media Data

Human behaviour is characterized by connections to others

Digital technologies have led to an explosion in the availability of networked data

EVOLUTION OF GROUPS

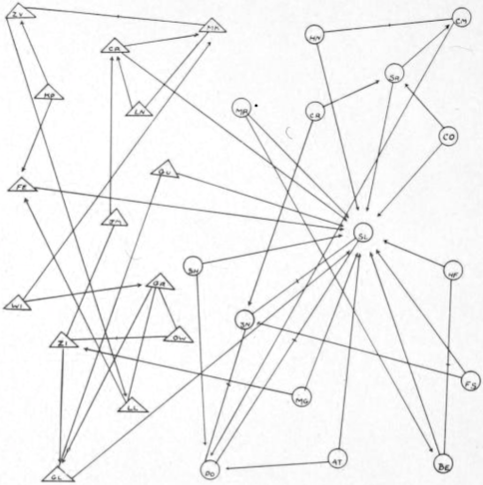


CLASS STRUCTURE, 1ST GRADE

21 boys and 14 girls. *Unchosen*, 18, GO, PR, CA, SH, FI, RS, DC, GA, SM, BB, TS, WI, KI, TA, HF, SA, SR, KR; *Pairs*, 3, EI-GO, WO-CE, CE-HN; *Stars*, 5, CE, WO, HC, FA, MB; *Chains*, 0; *Triangles*, 0; *Inter-sexual Attractions*, 22.

Moreno, “Who Shall Survive?” (1934)

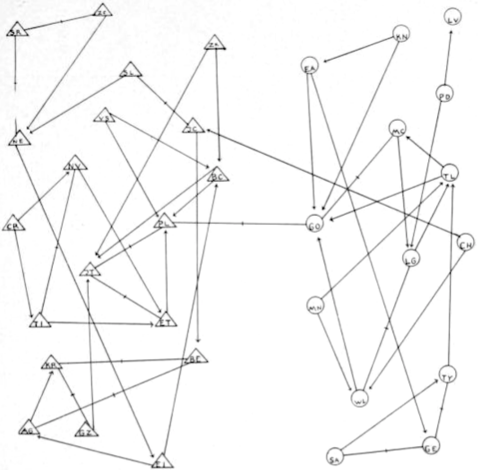
EVOLUTION OF GROUPS



CLASS STRUCTURE, 2ND GRADE

14 boys and 14 girls. *Unchosen*, 9, WI, KP, MG, AT, FS, CN, CR, MR, SH; *Pairs*, 11, ZV-MK, MK-LN, OW-ZI, GR-LL, ZI-JM, HN-CM, SL-JN, JN-PO, PO-SL, HF-BE, GL-GU; *Stars*, 2, SL, PO; *Chains*, 0; *Triangles*, 1, SL-JN-PO; *Inter-sexual Attractions*, 5.

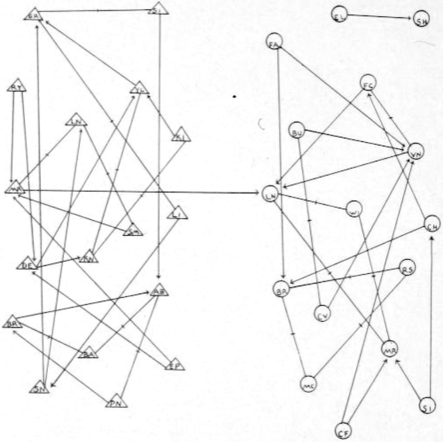
Moreno, “Who Shall Survive?” (1934)



CLASS STRUCTURE, 3RD GRADE

19 boys and 14 girls. *Unchosen,* **7**, VS, CR, CH, MN, PO, KN, ZK; *Pairs,* 14, SR-ZC, SR-NE, SL-JC, NV-TI, PL-JT, JT-ET, KR-BE, BE-AG, RR-GZ, PL-GO, GO-MC, WL-LG, SA-GE, GE-TY; *Stars,* **3**, GO, PL, JT; *Chains,* 1, ET-JT-PL-GO-MC; *Triangles,* **0**; *Inter-sexual Attractions,* **3**.

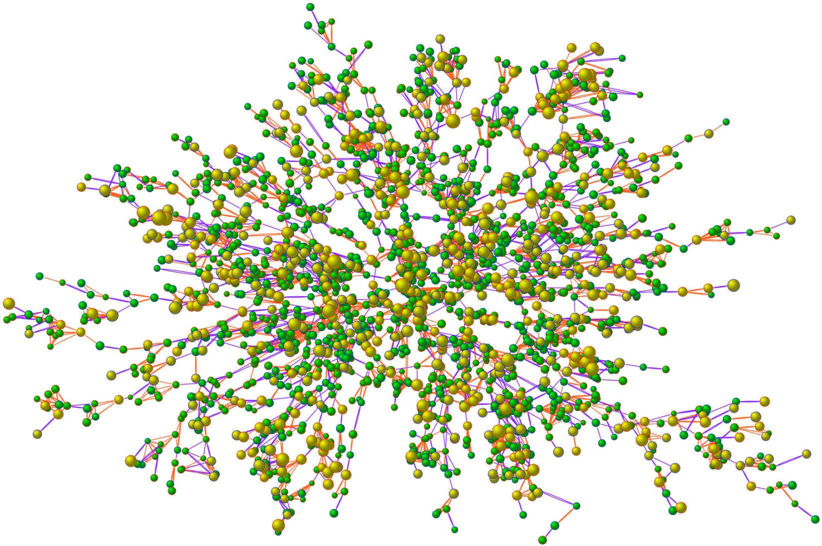
Moreno, "Who Shall Survive?" (1934)



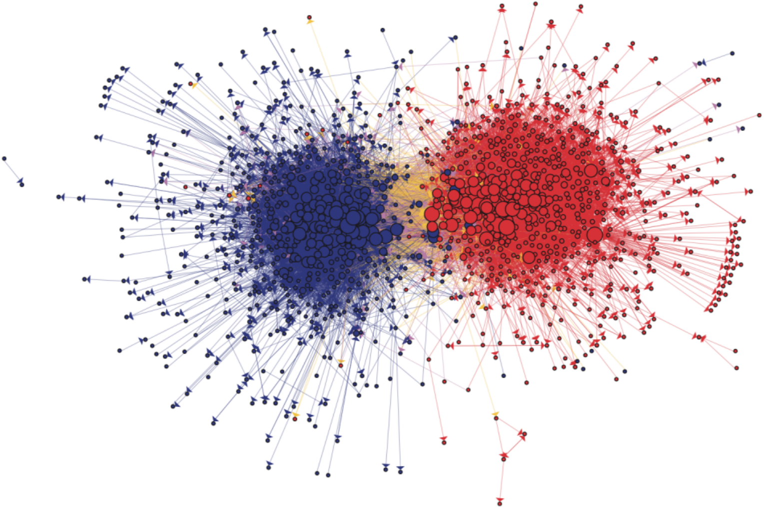
CLASS STRUCTURE, 4TH GRADE

17 boys and 16 girls. *Unchosen,* 6, EP, RY, EL, FA, SI, CF; *Pairs,* 17, GR-SI, GR-LI, MR-LN, LN-SM, YL-KN, AB-BA, BA-BR, KI-KN, AB-PN, FC-VN, BU-CV, LN-WI, LN-MR, BR-MC, BR-RS, WI-MR, MC-RS; *Stars,* 2, LN, VN; *Chains,* 0; *Triangles,* 2, BR-RS-MC; LN-WI-MR; *Inter-sexual Attractions,* 1.

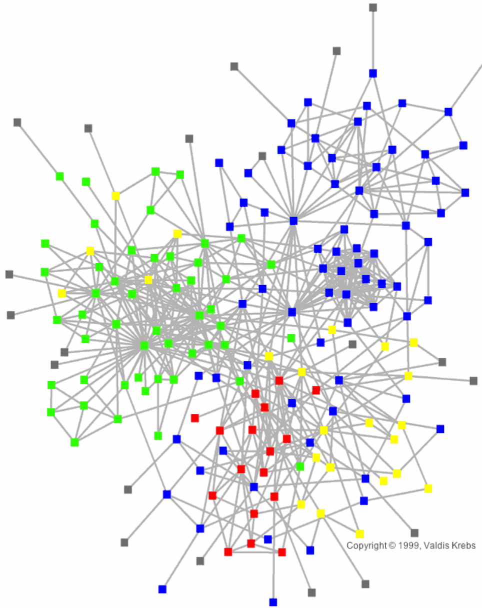
Moreno, “Who Shall Survive?” (1934)



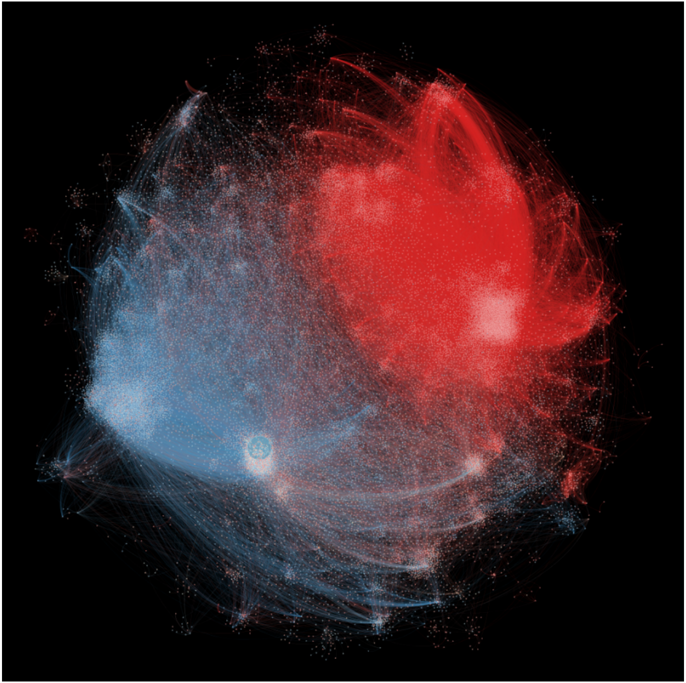
Christakis & Fowler, NEJM, 2007



Adamic & Glance, 2004, IWLD



Email network of a company



Barbera et al, 2015, Psychological Science

(Quick) introduction to social network analysis

What we will cover:

- Familiarity with **language of social network analysis**

(Quick) introduction to social network analysis

What we will cover:

- Familiarity with **language of social network analysis**
- Two key dimensions to analyze:

(Quick) introduction to social network analysis

What we will cover:

- Familiarity with **language of social network analysis**
- Two key dimensions to analyze:
  - **Centrality**: who is most influential in a network?

(Quick) introduction to social network analysis

What we will cover:

- Familiarity with **language of social network analysis**
- Two key dimensions to analyze:
  - **Centrality**: who is most influential in a network?
  - **Structure**: how to discover communities in a network?

(Quick) introduction to social network analysis

What we will cover:

- Familiarity with **language of social network analysis**
- Two key dimensions to analyze:
  - **Centrality**: who is most influential in a network?
  - **Structure**: how to discover communities in a network?
- Characteristics of networks that emerge in **digital environments**, such as social media sites

Basic concepts

- Node (vertex): each of the units in the network

Basic concepts

- Node (vertex): each of the units in the network
- Edge (tie): connection between nodes

Basic concepts

- Node (vertex): each of the units in the network
- Edge (tie): connection between nodes
  - Undirected: symmetric connection, represented by lines

Basic concepts

- Node (vertex): each of the units in the network
- Edge (tie): connection between nodes
  - Undirected: symmetric connection, represented by lines
  - Directed: imply direction, represented by arrows

Basic concepts

- Node (vertex): each of the units in the network
- Edge (tie): connection between nodes
  - Undirected: symmetric connection, represented by lines
  - Directed: imply direction, represented by arrows
  - Unweighted: all edges have same strength

Basic concepts

- Node (vertex): each of the units in the network
- Edge (tie): connection between nodes
  - Undirected: symmetric connection, represented by lines
  - Directed: imply direction, represented by arrows

  - Unweighted: all edges have same strength
  - Weighted: some edges have more strength than others

Basic concepts

- Node (vertex): each of the units in the network
- Edge (tie): connection between nodes
  - Undirected: symmetric connection, represented by lines
  - Directed: imply direction, represented by arrows

  - Unweighted: all edges have same strength
  - Weighted: some edges have more strength than others
- A network consists of a set of nodes and edges

Basic concepts

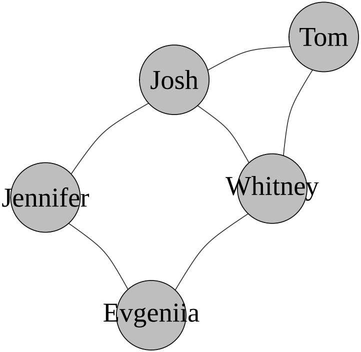
- Node (vertex): each of the units in the network
- Edge (tie): connection between nodes
  - Undirected: symmetric connection, represented by lines
  - Directed: imply direction, represented by arrows

  - Unweighted: all edges have same strength
  - Weighted: some edges have more strength than others
- A network consists of a set of nodes and edges

  i.e. a set of actors and their relationships

Basic concepts

Network Visualization

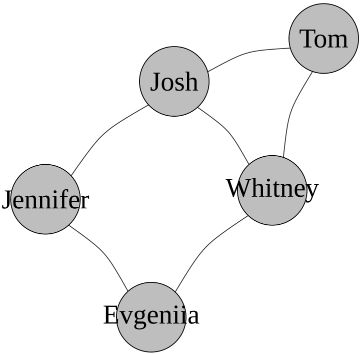


Adjacency Matrix

	P	J	E	W	T
P	0	1	1	0	0
J	1	0	0	1	1
E	1	0	0	1	0
W	0	1	1	0	1
T	0	1	0	1	0

Basic concepts

Network Visualization



Edgelist

	Node1	Node2
1	Paul	Josh
2	Paul	Evgeniia
3	Josh	Whitney
4	Josh	Tom
5	Whitney	Tom
6	Evgeniia	Whitney

Types of social media networks

- Internet: websites / hyperlinks

Types of social media networks

- Internet: websites / hyperlinks
- Twitter: users / retweets

Types of social media networks

- Internet: websites / hyperlinks
- Twitter: users / retweets
- Twitter: users / following connections

Types of social media networks

- Internet: websites / hyperlinks
- Twitter: users / retweets
- Twitter: users / following connections
- Twitter: hashtags / co-appeareance

Types of social media networks

- Internet: websites / hyperlinks
- Twitter: users / retweets
- Twitter: users / following connections
- Twitter: hashtags / co-appeareance
- Facebook: friends / friendship connections

Types of social media networks

- Internet: websites / hyperlinks
- Twitter: users / retweets
- Twitter: users / following connections
- Twitter: hashtags / co-appeareance
- Facebook: friends / friendship connections
- Reddit: subreddits / users in common

Social network analysis: key dimensions of analysis

Node centrality

How to measure actor influence or importance in a network?

Node centrality

How to measure actor influence or importance in a network?

Two main conceptual definition of centrality:

1. **Degree centrality**: number of connections for each node (potential for direct reach)

Node centrality

How to measure actor influence or importance in a network?

Two main conceptual definition of centrality:

1. **Degree centrality**: number of connections for each node (potential for direct reach)
   - Indegree: incoming connections

Node centrality

How to measure actor influence or importance in a network?

Two main conceptual definition of centrality:

1. **Degree centrality**: number of connections for each node (potential for direct reach)
    - Indegree: incoming connections
    - Outdegree: outgoing connections

Node centrality

How to measure actor influence or importance in a network?

Two main conceptual definition of centrality:

1. **Degree centrality**: number of connections for each node (potential for direct reach)
   - Indegree: incoming connections
   - Outdegree: outgoing connections
2. **Betweenness centrality**: gatekeeping potential

Node centrality

How to measure actor influence or importance in a network?

Two main conceptual definition of centrality:

1. **Degree centrality**: number of connections for each node (potential for direct reach)
   - Indegree: incoming connections
   - Outdegree: outgoing connections
2. **Betweenness centrality**: gatekeeping potential
   - How well a node connects different parts of the network

Node centrality

How to measure actor influence or importance in a network?

Two main conceptual definition of centrality:

1. **Degree centrality**: number of connections for each node (potential for direct reach)
   - Indegree: incoming connections
   - Outdegree: outgoing connections
2. **Betweenness centrality**: gatekeeping potential
   - How well a node connects different parts of the network
   - Fraction of shortest paths between any two nodes on which a particular node lies

Node centrality

How to measure actor influence or importance in a network?

Two main conceptual definition of centrality:

1. **Degree centrality**: number of connections for each node (potential for direct reach)
   - Indegree: incoming connections
   - Outdegree: outgoing connections
2. **Betweenness centrality**: gatekeeping potential
   - How well a node connects different parts of the network
   - Fraction of shortest paths between any two nodes on which a particular node lies

$\rightarrow$ Other measures:

Node centrality

How to measure actor influence or importance in a network?

Two main conceptual definition of centrality:

1. **Degree centrality**: number of connections for each node (potential for direct reach)
    - Indegree: incoming connections
    - Outdegree: outgoing connections
2. **Betweenness centrality**: gatekeeping potential
    - How well a node connects different parts of the network
    - Fraction of shortest paths between any two nodes on which a particular node lies

$\rightarrow$ Other measures:

- **Closeness centrality**: broadcasting potential

Node centrality

How to measure actor influence or importance in a network?

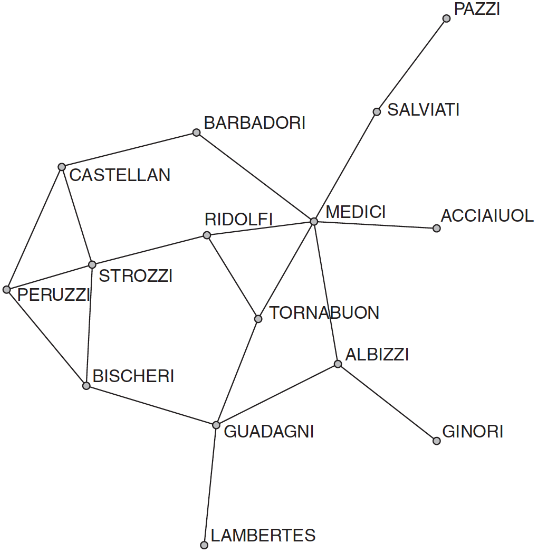
Two main conceptual definition of centrality:

1. **Degree centrality**: number of connections for each node (potential for direct reach)
   - Indegree: incoming connections
   - Outdegree: outgoing connections
2. **Betweenness centrality**: gatekeeping potential
   - How well a node connects different parts of the network
   - Fraction of shortest paths between any two nodes on which a particular node lies

$\rightarrow$ Other measures:

- **Closeness centrality**: broadcasting potential
- **Eigenvector centrality and coreness**: centrality measured as being connected to other central neighbors

Florentine family marriages in the 15th century

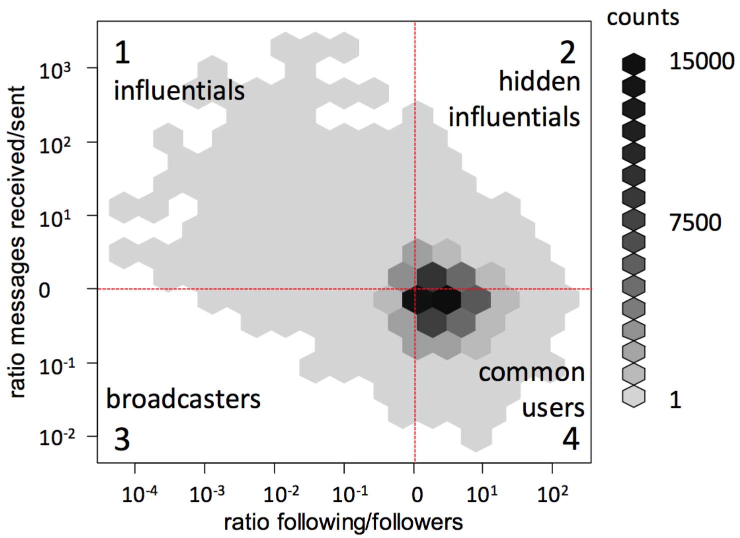


Source: Padgett (1993) and Sinclair (2016)

Occupy Wall Street Twitter networks

Source: Lotan (2011)

Protest networks on Twitter



Source: González-Bailón et al (2013)

Occupy Wall Street Twitter networks

Information Brokers

RT @GlobalRevLive: #TocandoaBankia live spanishrevolution.tv #mayo2013 fb.me/2Q6jOlkUs #ows #stopdeshuacios #spanishrevolution
Expand

N=8,082

N=51,212

N=74,007

Source: González-Bailón and Wang (2016)

Discovery in large-scale networks

How to understand the structure of large-scale networks?

- Latent communities or clusters

Discovery in large-scale networks

How to understand the structure of large-scale networks?

- Latent communities or clusters
    - **Community detection algorithms**

Discovery in large-scale networks

How to understand the structure of large-scale networks?

- Latent communities or clusters
  - **Community detection algorithms**
  - Finding groups of nodes that densely connected internally, more so than to the rest of the networks

Discovery in large-scale networks

How to understand the structure of large-scale networks?

- Latent communities or clusters
  - **Community detection algorithms**
  - Finding groups of nodes that densely connected internally, more so than to the rest of the networks
  - Overlap with shared visible or latent similarities (homophily)

Discovery in large-scale networks

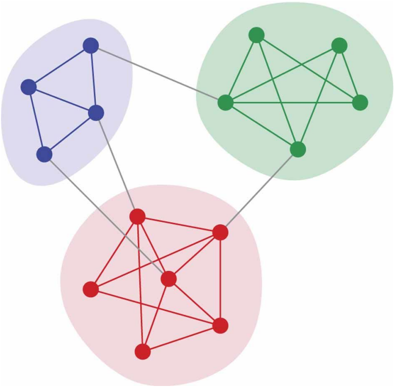
How to understand the structure of large-scale networks?

- Latent communities or clusters
  - **Community detection algorithms**
  - Finding groups of nodes that densely connected internally, more so than to the rest of the networks
  - Overlap with shared visible or latent similarities (homophily)
  - Also hierarchy: core-periphery detection

Community detection

Community structure:

- Network nodes often cluster into tightly-knit groups with a high density of within-group edges and a lower density of between-group edges
- **Modularity score**: measures clustering of nodes compared to random network of same size
- Many different community detection algorithms based on different assumptions



Source: Newman (2012)

Network hierarchy

- **Intuition**

Network hierarchy

- **Intuition**
  - Large-scale networks have hierarchical properties

Network hierarchy

- **Intuition**
  - Large-scale networks have hierarchical properties
- **Network core:**

Network hierarchy

- **Intuition**
  - Large-scale networks have hierarchical properties
- **Network core:**
  1. *Centrality*: high relative importance in network

Network hierarchy

- **Intuition**
  - Large-scale networks have hierarchical properties
- **Network core:**
  1. *Centrality*: high relative importance in network
  2. *Connectivity*: many possible distinct paths between individuals

Network hierarchy

- **Intuition**
  - Large-scale networks have hierarchical properties
- **Network core:**
  1. *Centrality*: high relative importance in network
  2. *Connectivity*: many possible distinct paths between individuals
     (not captured by simple topological measures)

Network hierarchy

- **Intuition**
  - Large-scale networks have hierarchical properties
- **Network core:**
  1. *Centrality*: high relative importance in network
  2. *Connectivity*: many possible distinct paths between individuals
     (not captured by simple topological measures)
- **k-core decomposition**

Network hierarchy

- **Intuition**
  - Large-scale networks have hierarchical properties
- **Network core:**
  1. *Centrality*: high relative importance in network
  2. *Connectivity*: many possible distinct paths between individuals
     (not captured by simple topological measures)
- **k-core decomposition**
  - Algorithm to partition a network in nested shells of connectivity

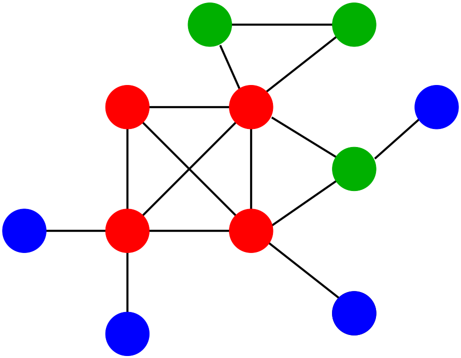
Network hierarchy

- **Intuition**
  - Large-scale networks have hierarchical properties
- **Network core:**
  1. *Centrality*: high relative importance in network
  2. *Connectivity*: many possible distinct paths between individuals
     (not captured by simple topological measures)
- **k-core decomposition**
  - Algorithm to partition a network in nested shells of connectivity
  - The k-core of a graph is the maximal subgraph in which every node has at least degree k

Network hierarchy

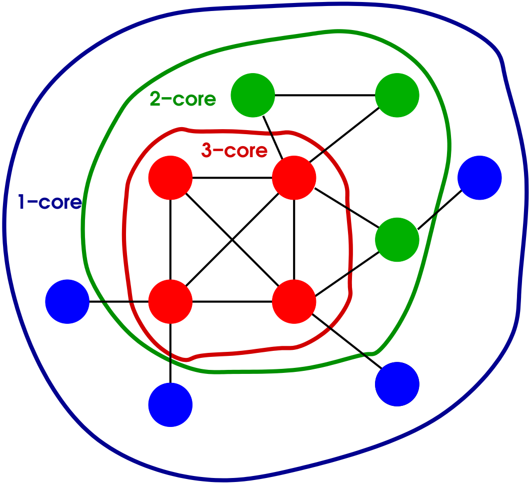
- **Intuition**
  - Large-scale networks have hierarchical properties
- **Network core:**
  1. *Centrality*: high relative importance in network
  2. *Connectivity*: many possible distinct paths between individuals
     (not captured by simple topological measures)
- **k-core decomposition**
  - Algorithm to partition a network in nested shells of connectivity
  - The k-core of a graph is the maximal subgraph in which every node has at least degree k
  - Many applications; scales well to large networks.

k-core decomposition



Source: Alvarez-Hamelin et al, 2005

k-core decomposition



Source: Alvarez-Hamelin et al, 2005

k-core decomposition of #OccupyGezi network