The 8th International Conference on Network Analysis

Moscow, Russia, May 18-19, 2018

Machine Learning Analysis of Complex Networks in Hyperspherical Space

Ernesto Estrada

ernesto.estrada@strath.ac.uk

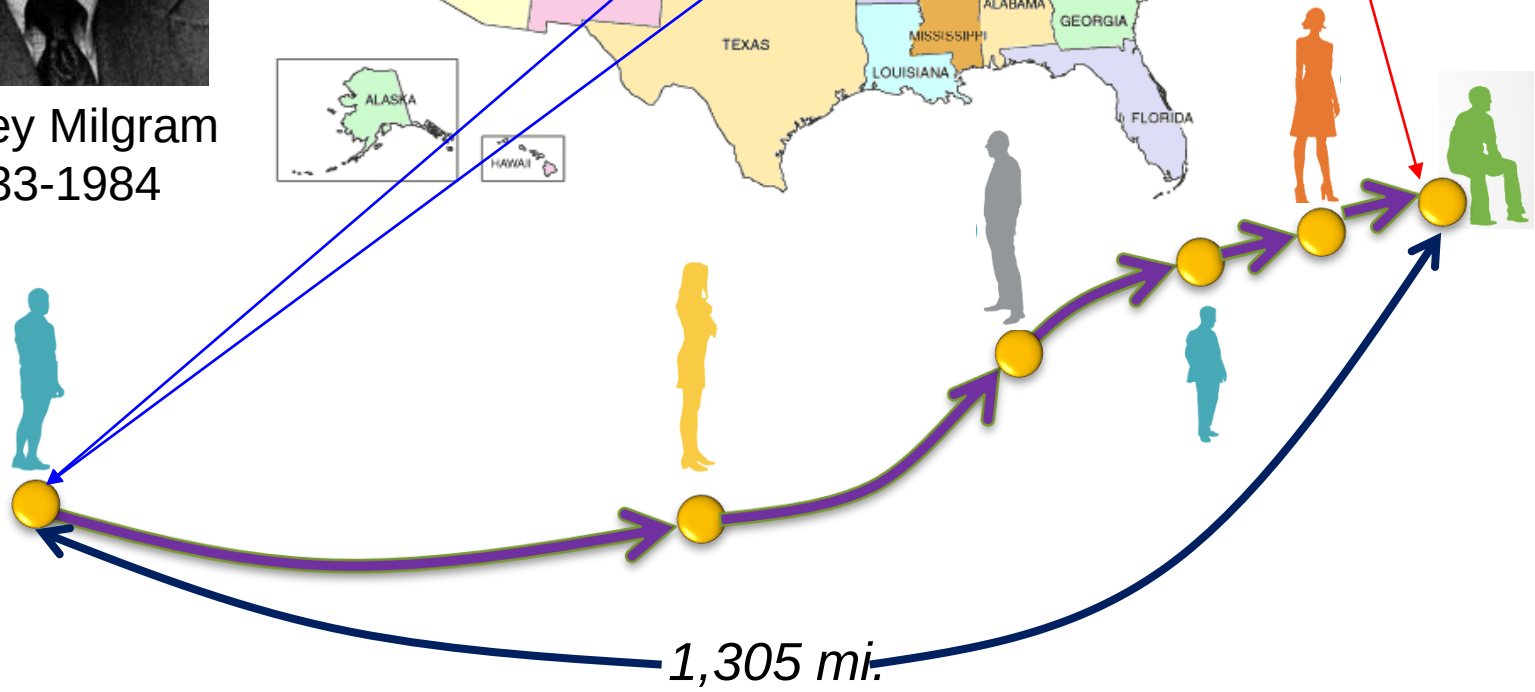
www.estradalab.org

@eestradalab

Motivation

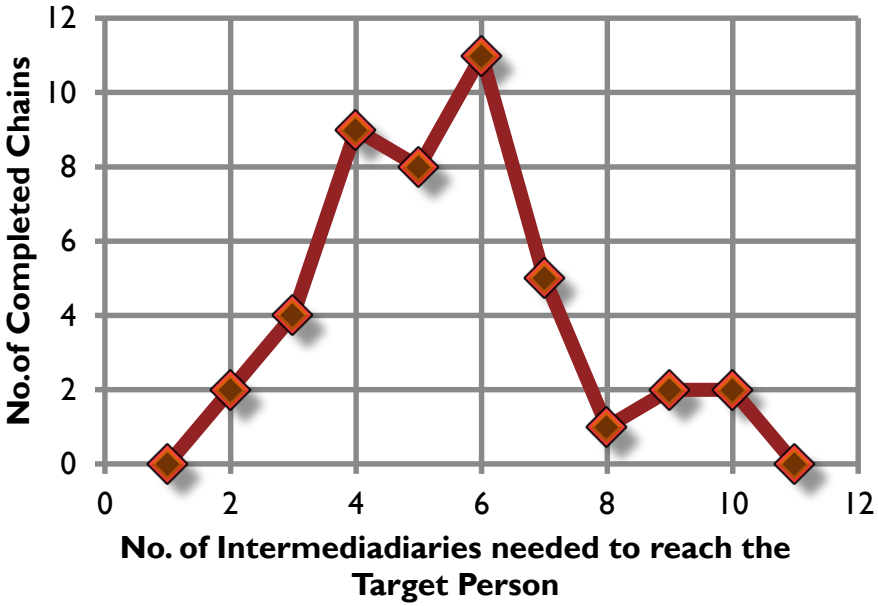
The 'Small-World' Phenomenon

Stanley Milgram
1933-1984



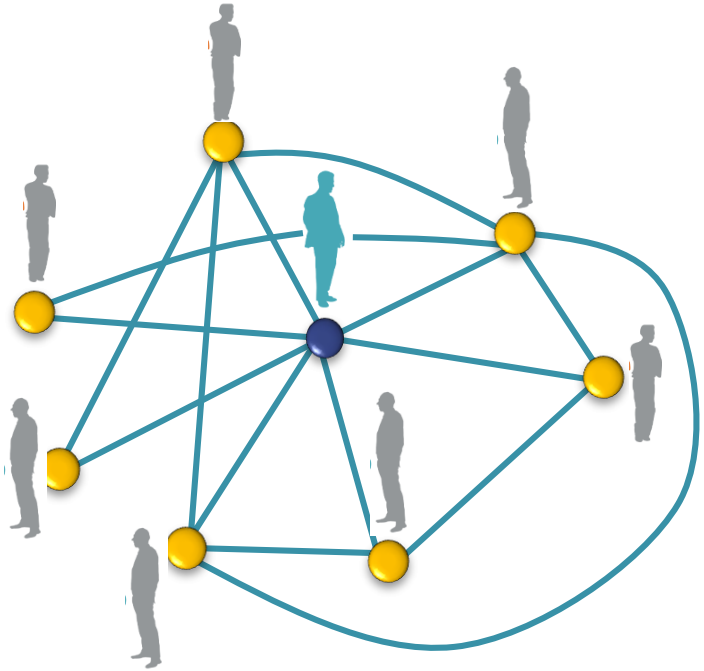
'Small-World' Phenomenon

Notion of Distance



Accounts for the successful transfer of information between the senders and the target

Notion of Transitivity

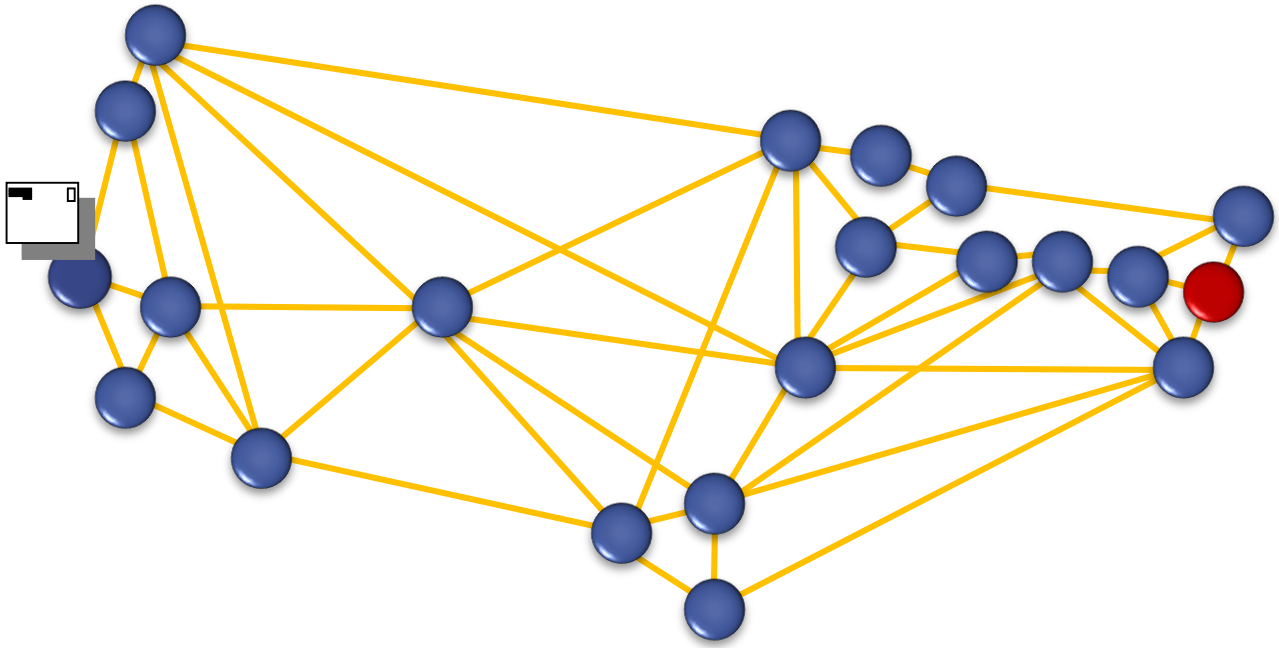


Accounts for part of the information lost in the communication between the senders and the target

Milgram, S., 1967. *Psychology today*, *2*(1), pp.60-67.

Shortest Path Distance

Definition: A path of length l is a sequence of distinct nodes $v_1, v_2, \ldots, v_l$ such as for each $i=1,2,\ldots,l$ there is an edge from v_i to v_{i+1}.



$$d_{ij} = 5$$

Shortest Path Distance

Network Global Efficiency

$$E(G) = \frac{1}{n(n-1)} \sum_{i<j} \frac{1}{d_{ij}}$$

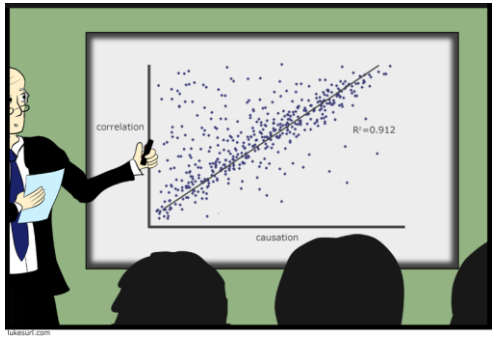
Definition of "efficiency": the ratio of the work done or energy developed by a machine, engine, etc., to the energy supplied to it, usually expressed as a percentage.

EFFICIENCY FOR WHAT?

Latora & Marchiori. Phys. Rev. Lett. ***87*** *(2001) 198171*

Shortest Path Approaches

Observation: Most real-world networks display relatively small average shortest path length: "small-world" phenomenon.

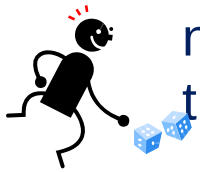


Inference: "Items" must travel mainly through the shortest paths of the networks.

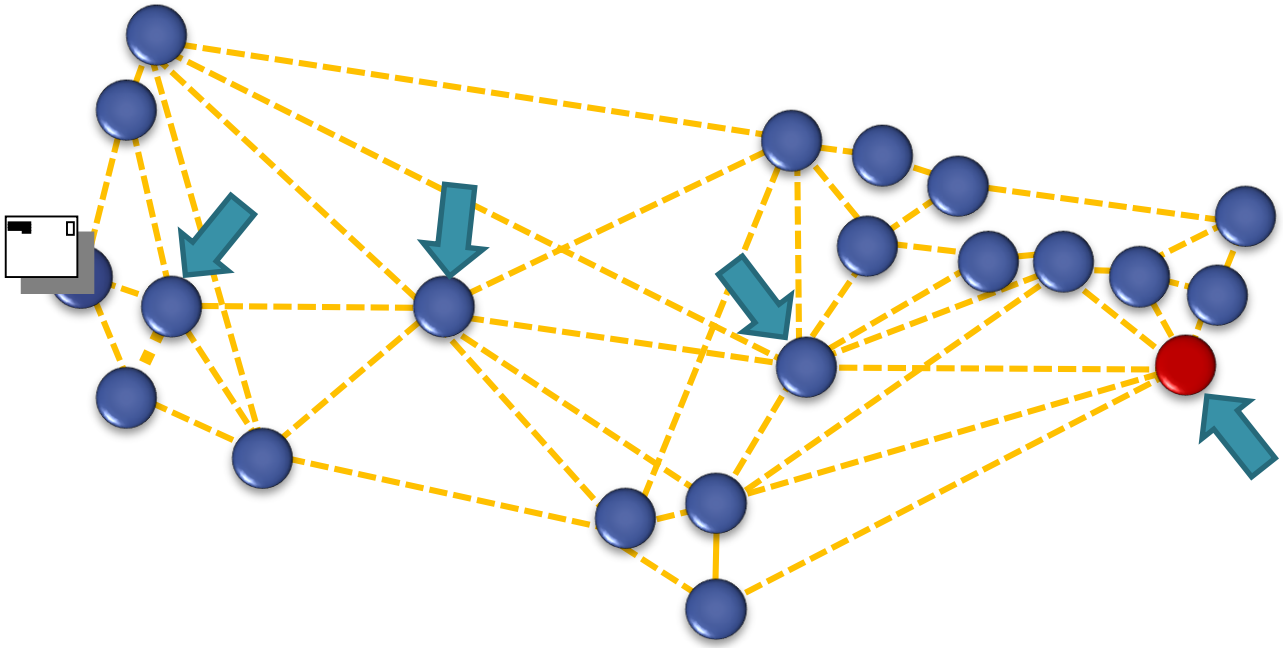
The troubles

Shortest Path Approaches

The sender does not know the global structure of the network!

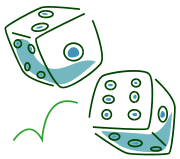


There is a high probability that the shortest path connecting nodes *p* and *q* goes through the most connected nodes of the network.

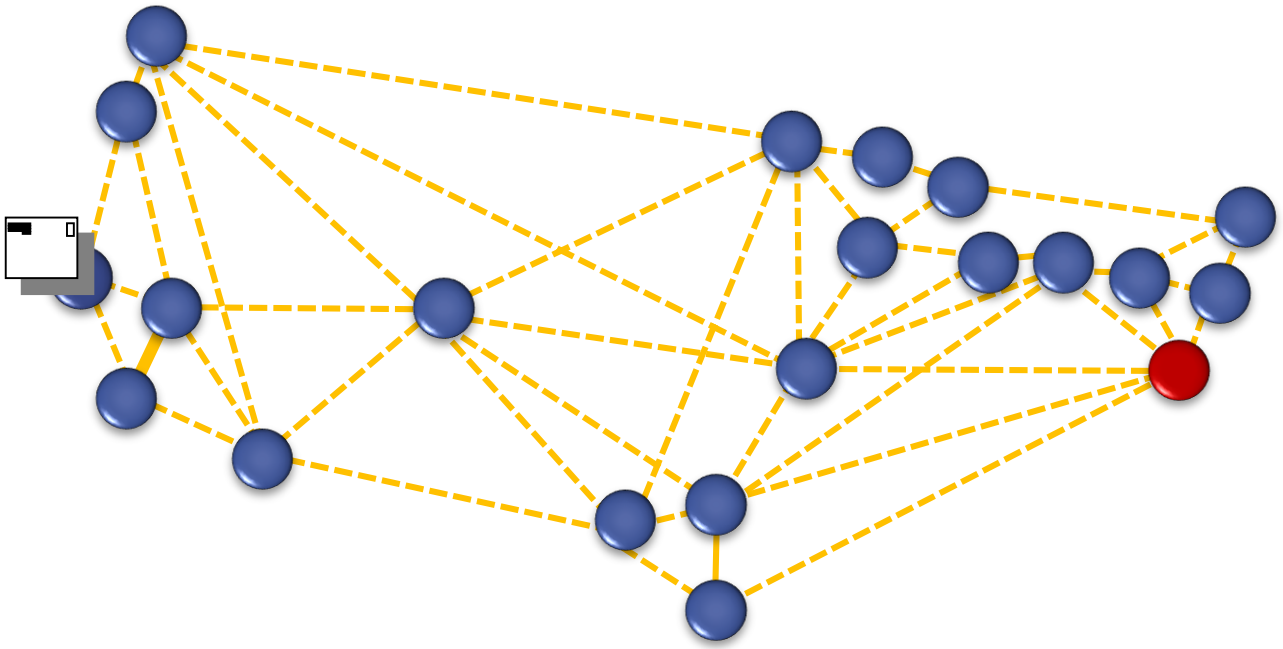


Sending the ‘items’ to the most connected nodes (hubs) of the network will increase the chances of reaching the target.

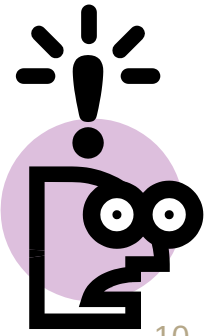
Shortest Path Approaches



But also, there is a high probability that the most connected nodes of the network are involved in a high number of transitive relations.



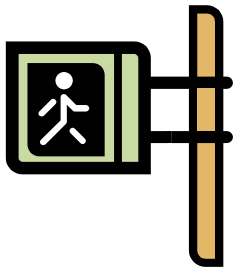
Sending ‘items’ to the most connected nodes (hubs) of the network will increase the chances of getting lost.



All-Routes Approaches

Hypothesis: The items not only flows through the shortest paths but by mean of any available route.

Definition: The communicability between two nodes in a network is defined as a function of the total *number of routes* connecting them, giving more importance to the shorter than to the longer ones.



Definition: A route is a walk of length l, which is any sequence of (not necessarily different) nodes $v_1,\ldots, v_l$ such as for each $i = 1,\ldots,l$ there is a link from v_l to v_{l+1}.

Estrada & Rodríguez-Velázquez: *Phys. Rev. E* 71, **2005**, 056103
Estrada & Hatano: *Phys. Rev. E* 77, **2008**, 036111

All-Routes Approaches

Communicability

The communicability function between the nodes p and q is then mathematically defined as

$$G_{pq} = \sum_{l=0}^{\infty} c_l \left(\#\ \text{of walks in}\ l\ \text{steps from}\ p\ \text{to}\ q \right)$$

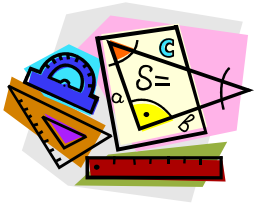
where c_l should:

- makes the series convergent
- gives more weight to the shorter than to the longer walks

Estrada & Rodríguez-Velázquez: *Phys. Rev. E* 71, **2005**, 056103
Estrada & Hatano: *Phys. Rev. E* 77, **2008**, 036111

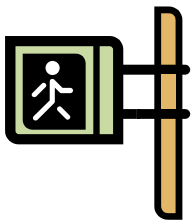
All-Routes Approaches

Communicability



Definition: Let $\ell_2(V) = \left\{ \{f(p)\}_{p\in V} \left| \sum_{p\in V} |f(p)|^2 < \infty \right. \right\}$.

The adjacency operator is a bounded operator on $\ell_2(V)$ defined as $(Af)(p) = \sum_{(p,q)\in E} f(q)$.



Theorem: The number of walks of length *l* between the nodes *p* and *q* in a network is equal to:

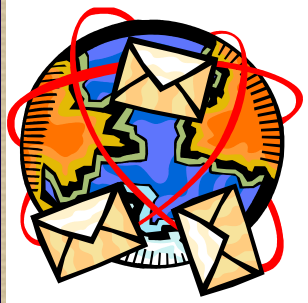
$\left(A^l\right)_{pq}$.

Estrada & Knight: *A First Course in Network Theory* **2015**, (Oxford Univ. Press).

All-Routes Approaches

Communicability

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be a nonincreasing ordering of the eigenvalues of *A*, and let $\vec{\psi}_j$ be the eigenvector associated with λ_j .



The *communicability function* can be written as:

$$G_{pq} = \sum_{l=0}^{\infty} c_l \left(A^l\right)_{pq} = \sum_{l=0}^{\infty} \sum_{j=1}^{n} c_l \lambda_j^l \psi_{j,p} \psi_{j,q}$$

Estrada & Rodríguez-Velázquez: *Phys. Rev. E* 71, **2005**, 056103
Estrada & Hatano: *Phys. Rev. E* 77, **2008**, 036111

All-Routes Approaches

Communicability function

$$G_{pq} = \sum_{l=0}^{\infty} \frac{\left(A^l\right)_{pq}}{l!}$$

$$= \left(e^A\right)_{pq}$$

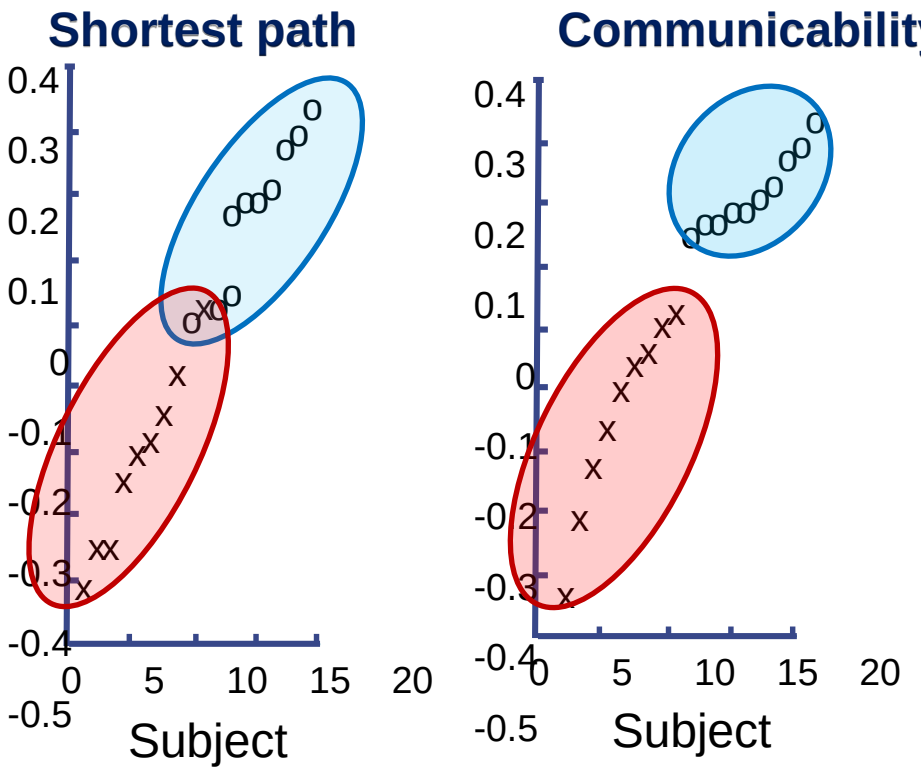
$$= \sum_{j=1}^{n} \psi_{j,p} \psi_{j,q} e^{\lambda_j}$$

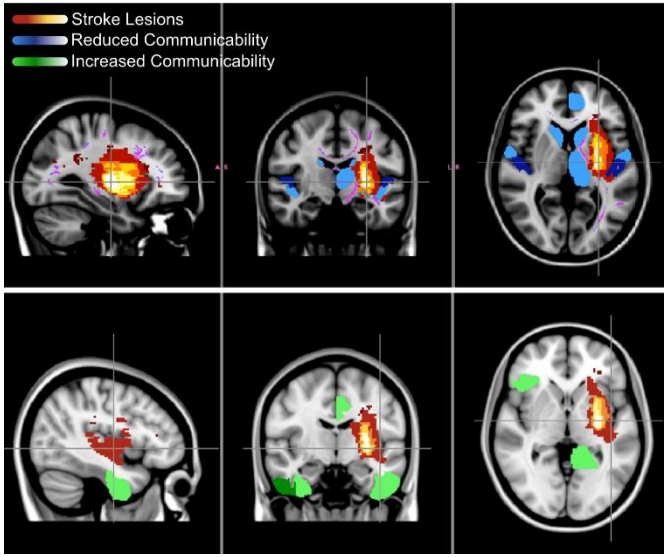
Estrada, Hatano & Benzi, *Phys. Rep*. 514 **2012**, 89-119
Estrada & Higham, *SIAM Rev*. 52, **2010**, 696-714
Estrada: *J. Theor. Biol*. 263 **2010**, 556-565.

Evidences

Experimental evidences

Communicability & brain diseases

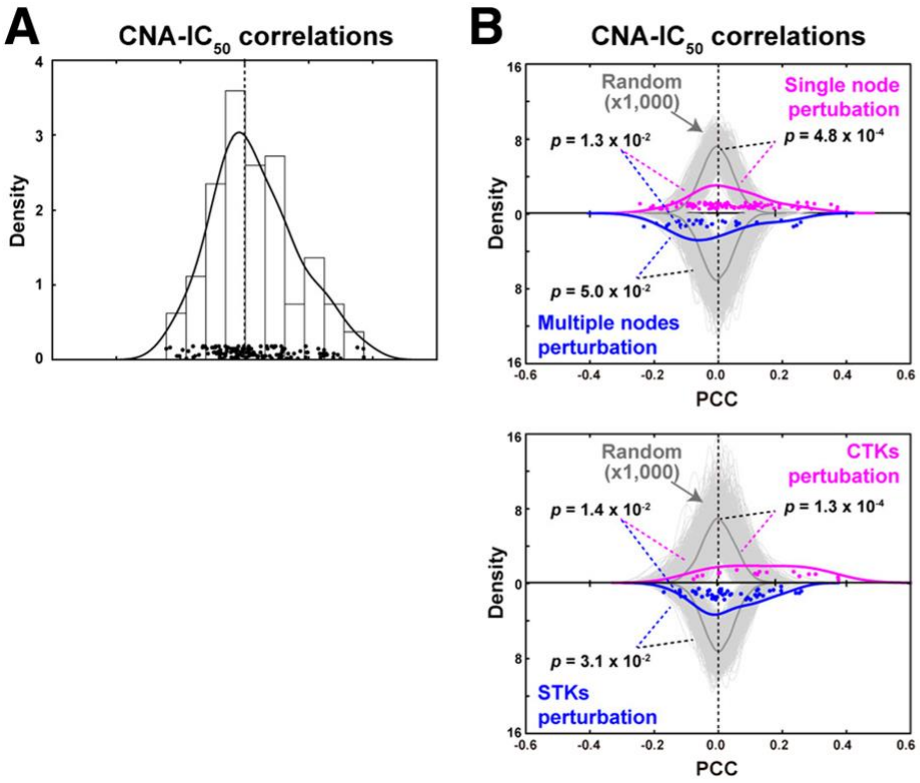




Crofts & Higham: *J. Roy. Soc. Interface* **6** (2009) 411
Crofts et al: *NeuroImage* **54** (2011) 161-9

Experimental evidences

Communicability & cancer therapeutics



C

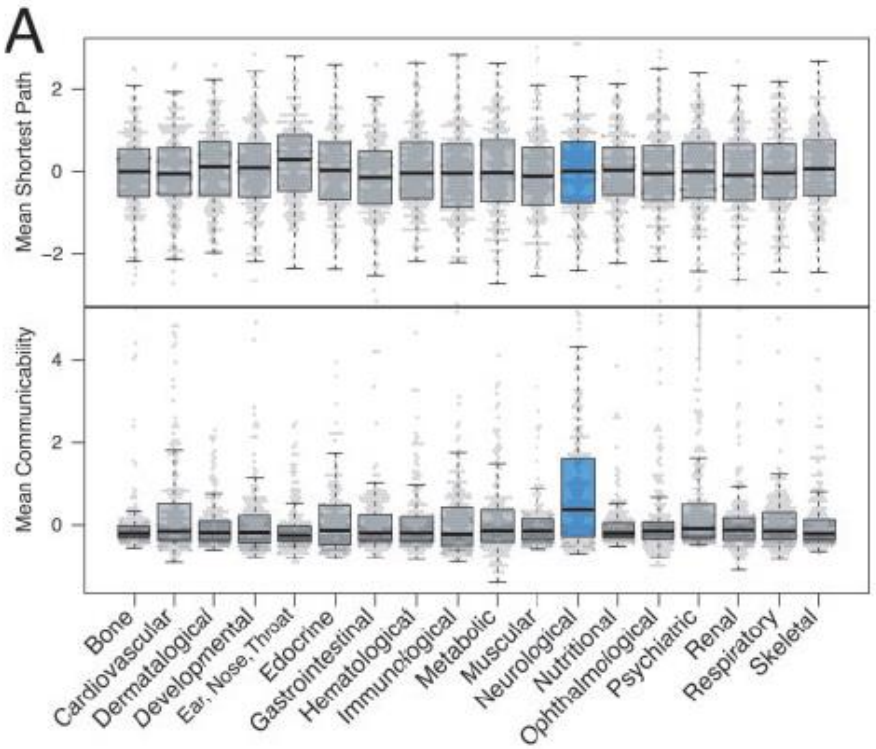
CNA-based differential associations ($p < 0.05$)	
Compound Class I	
Clinical	Experimental
Experimental	Chemotherapy
Experimental	In clinical development
Compound Class II	
Cytotoxic	Targeted
Target Family	
Chemotherapy	Receptor Tyrosine Kinase
Chemotherapy	CTK
Chemotherapy	STKs
STKs	CTKs
Effector Pathway	
DNA repair	Cytoskeleton
DNA repair	ERK signaling
ERK signaling	Apoptosis
Mitosis	DNA repair
PI3K/mTOR	DNA repair
Replication	Cytoskeleton
Replication	ERK ignaling
Replication	Mitosis
Replication	PI3K/mTOR
Stress pathway	Cytoskeleton
Stress pathway	ERK signaling
Stress pathway	Mitosis
Stress pathway	PI3K/mTOR
Transcription	Cytoskeleton
Transcription	ERK signaling
Transcription	Mitosis
Transcription	PI3K/mTOR

Cancer therapeutics response can be partially predicted on the basis of a network communicability measure that integrates gene expression and protein interaction data.

Serra-Musach et al. *Genome Medicine* (2016) 8:88

Experimental evidences

Communicability & human genetic diseases



Network communicability provides advantages over alternative metrics because it retains topology information, lends itself to set-based analysis and it is easy to represent with univariate scores. It outperforms shortest path in a variety of situations.

Campbell et al: *Bioinformatics* **30** (2014) 3387-3389

Experimental evidences

Communicability & autonomous vehicles

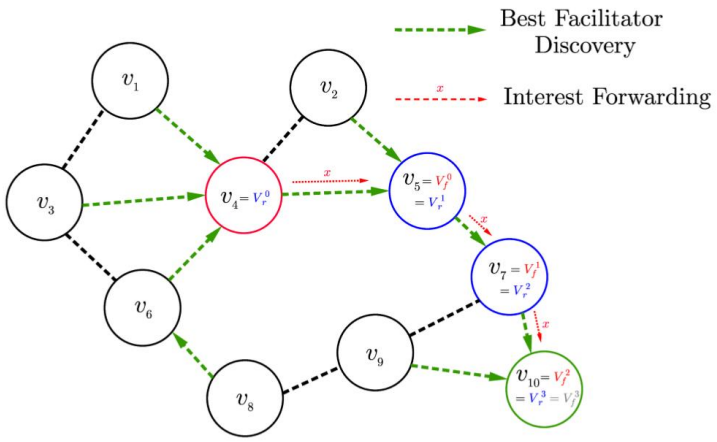


Figure 1: Facilitator Discovery Process

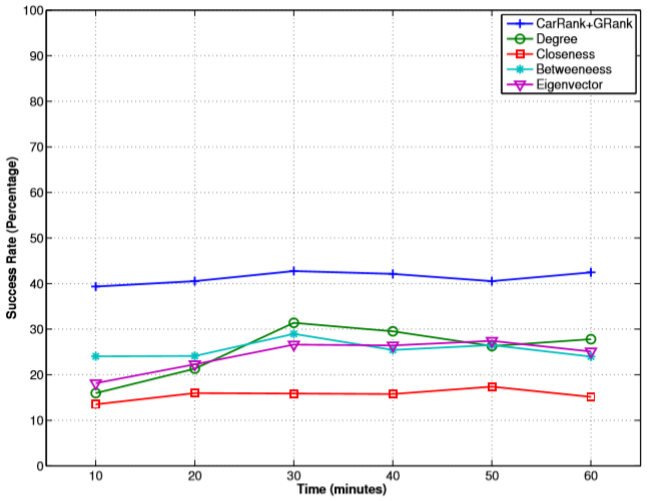


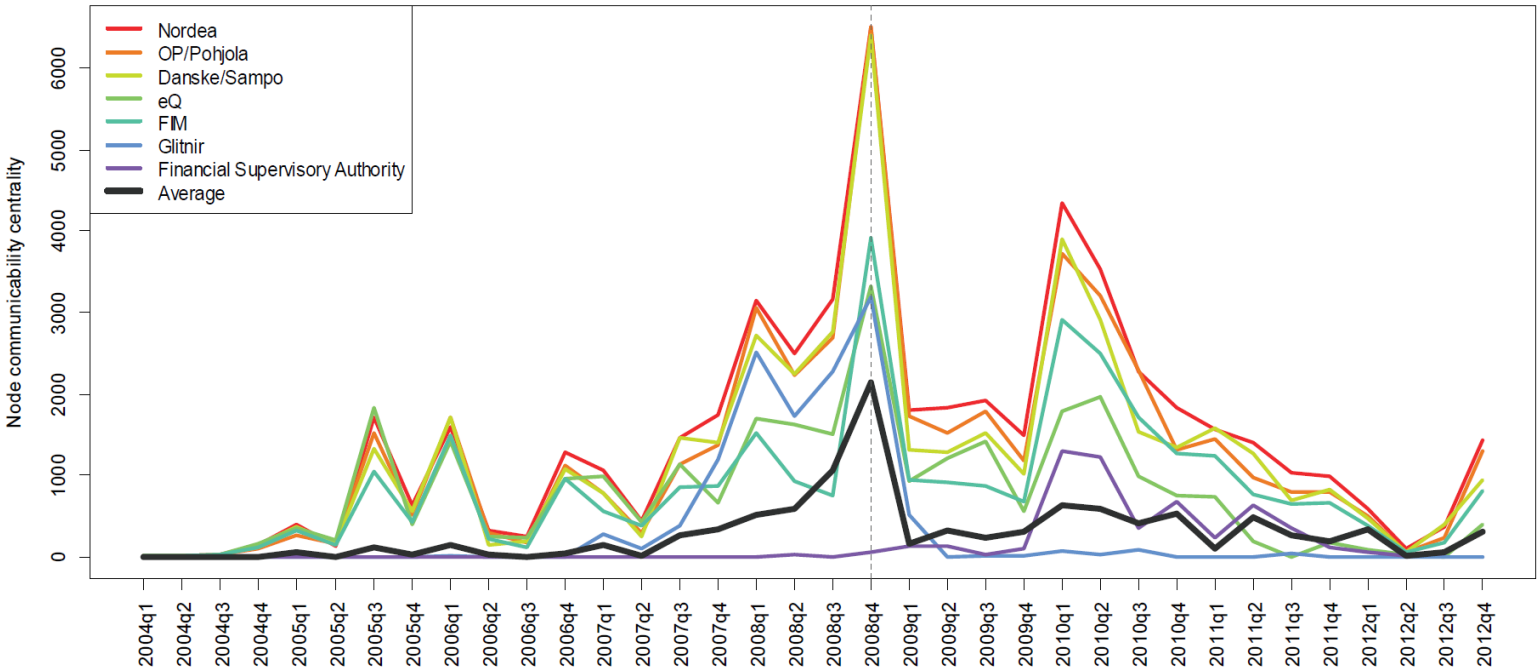
Figure 3: Success rate conparison for the consumer interests satisfied over time using different centrality schemes for content distribution

Inspired from the concept of communicability in complex networks, Grank, a global vehicle scheme allows a vehicles to use a new stable metric named "Information communicability" to rank different locations in the city and rank itself accordingly.

Khan, and Ghamri-Doudane. *IEEE Comm. Mag.* 54.8 (2016): 100-107.

Experimental evidences

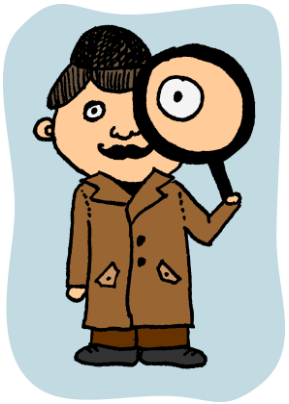
Communicability & financial crisis



An even clearer reaction is visible in network communicability. A second peak occurs around 2010Q1, primarily in communicability, which might be indirectly linked to revelations concerning the financial situation in Greece and its actively-debated first bailout.

Rönnqvist & Sarlin. In 2104 IEEE Conference Computational Intelligence for Financial Engineering & Economics (CIFEr).

Conclusions



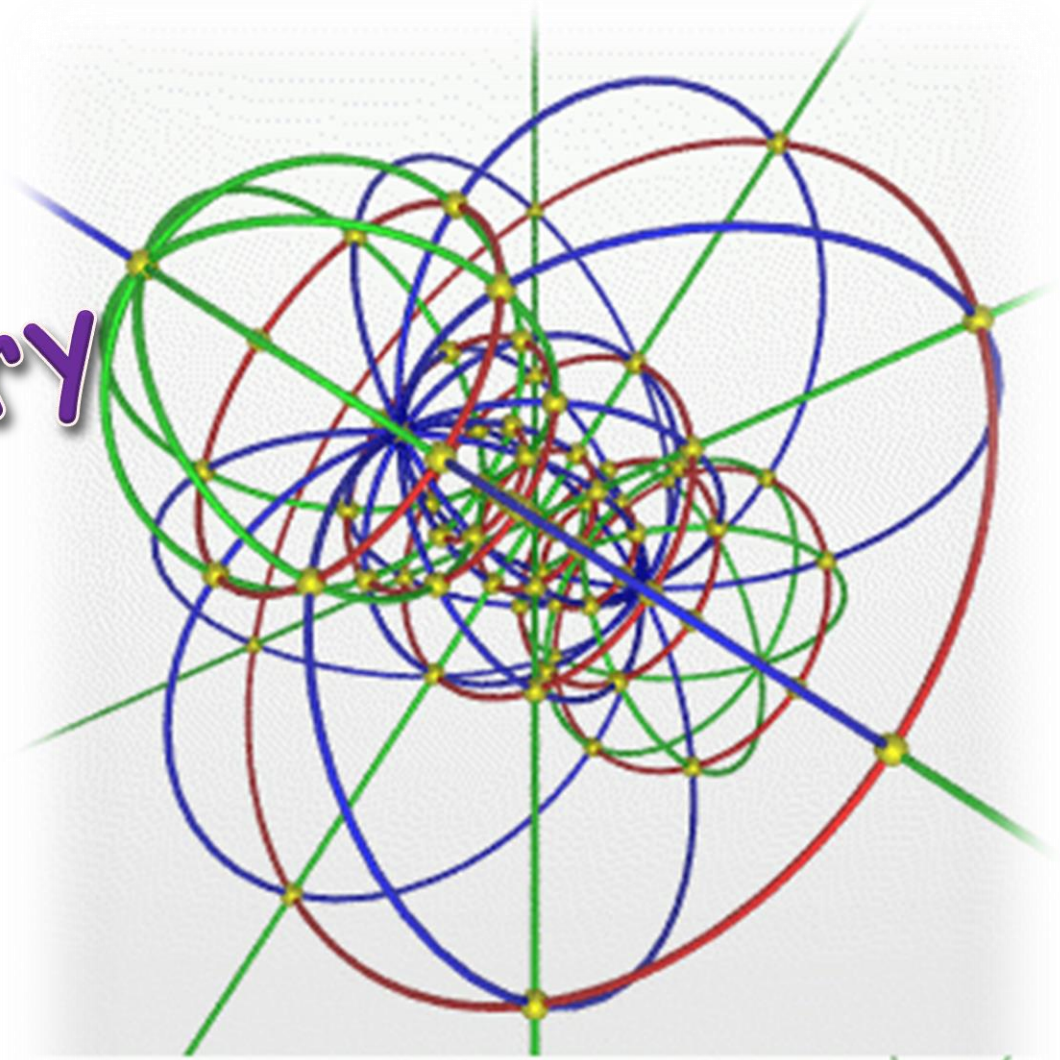
Observation: In many real-world processes on networks, communicability is more explicative than shortest path routes.

Inferences:

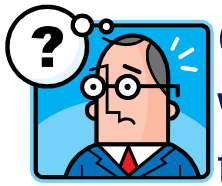
1. “Items” may travel in a diffusion-like way through the nodes and edges of networks.

2. “Items” may travel using an “all-routes” way more than through shortest paths only.

The Geometry



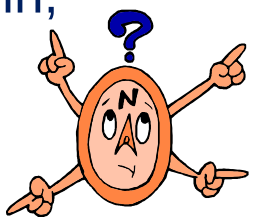
Quality of information



Consider that two nodes *p* and *q* are trying to communicate with each other by sending information both ways through the network.

We know that:

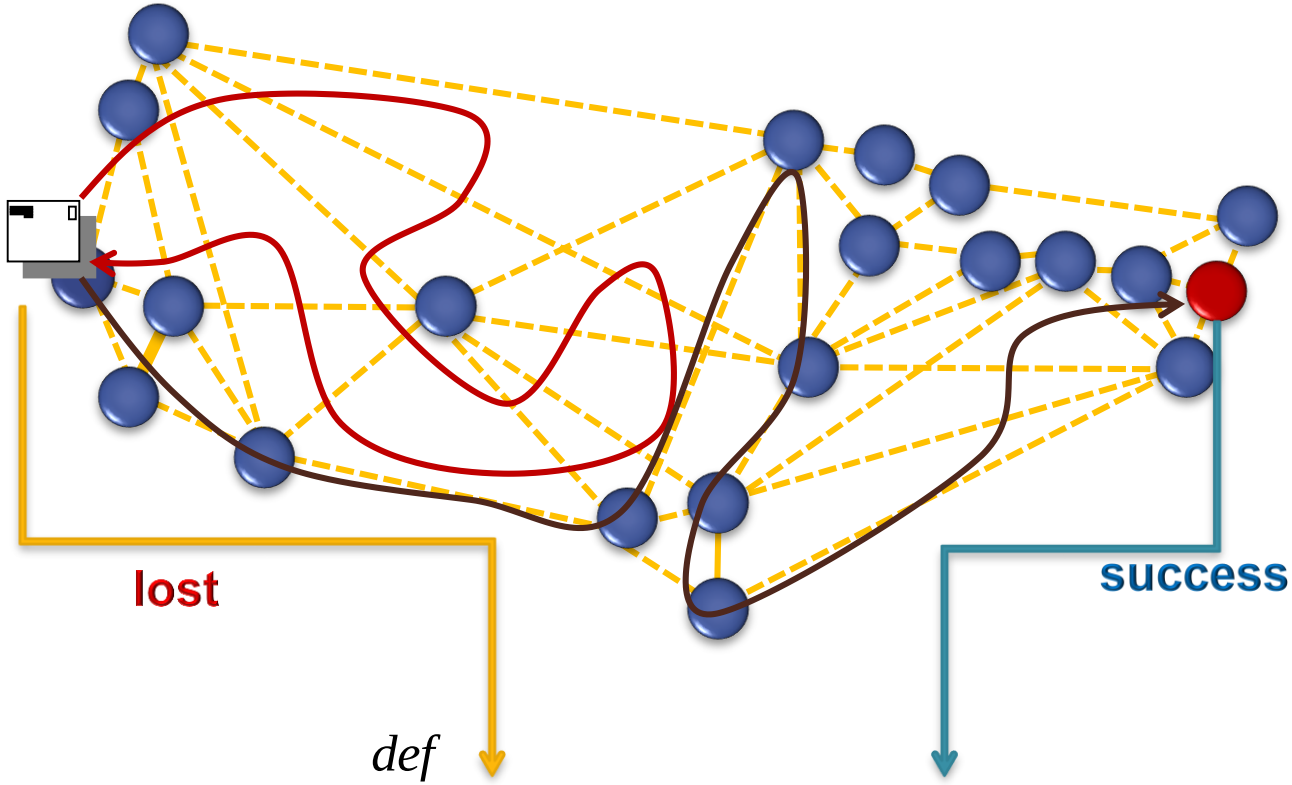
G_{pp} Quantifies the amount of 'information' that is sent by *p*, wanders around the network, and returns to the origin, the node *p*.



G_{pq} Quantifies the amount of 'information' that is sent by *p*, wanders around the network, and arrives to its destination, the node *q*.

Quality of information

The goal of communication is to maximise the amount of information that arrives to its destination by minimising that which is lost.

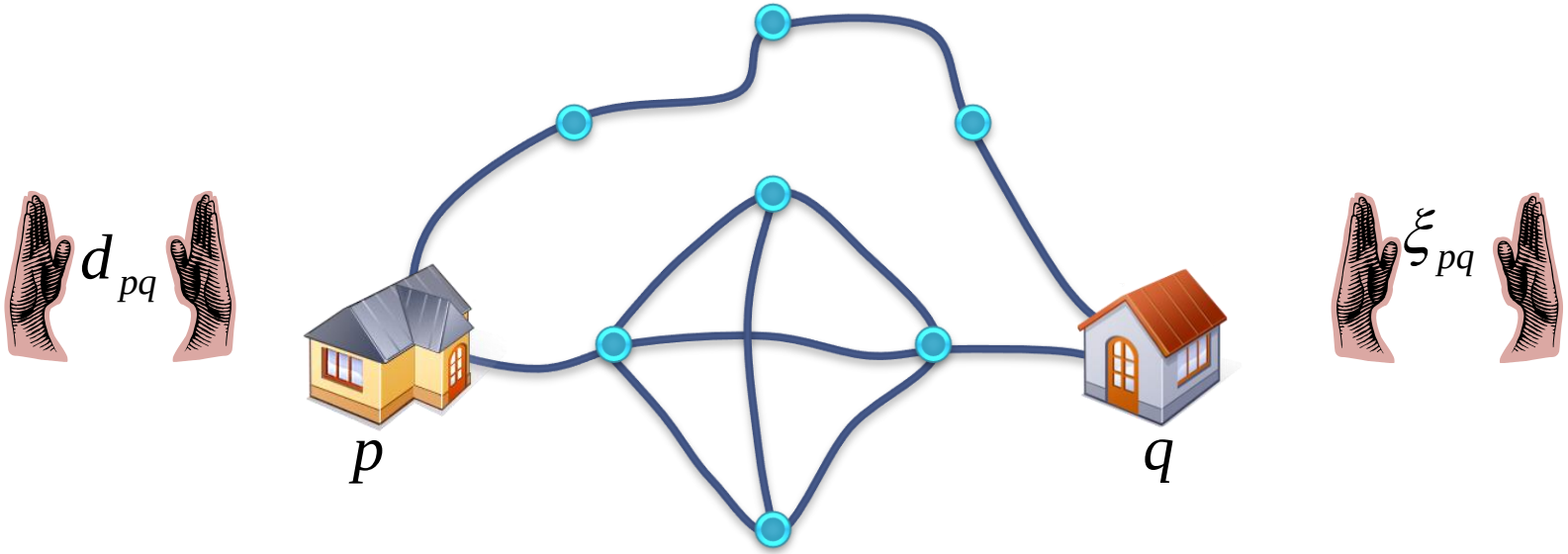


$$\xi_{pq} \overset{def}{=} G_{pp} + G_{qq} - 2G_{pq}$$

Estrada: *Lin. Algebra Appl.* 436 **2012** 4317-4328

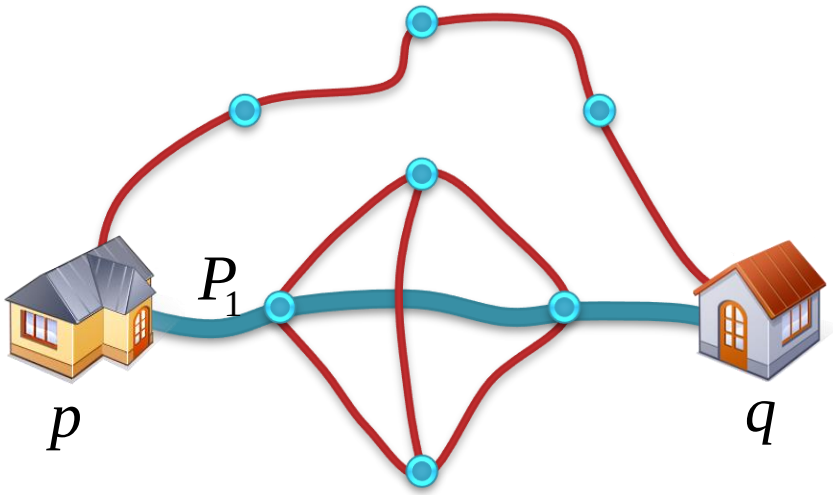
Communicability Distance

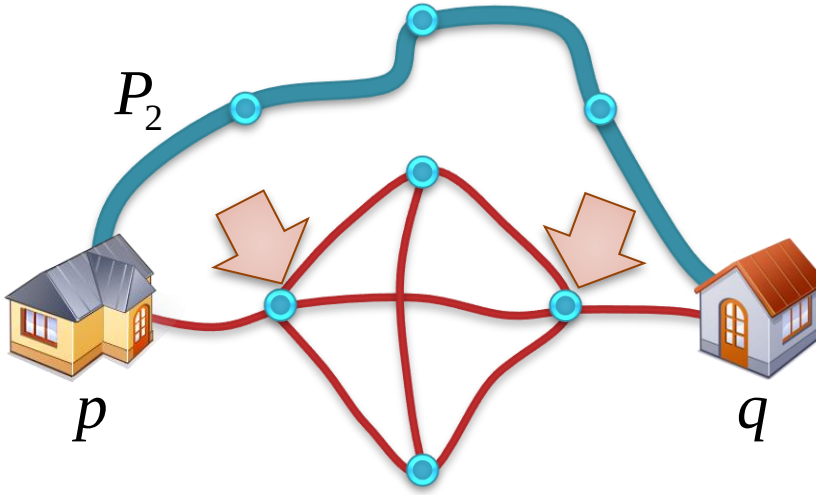
Theorem: The function ξ_{pq} is a Euclidean distance between the corresponding nodes.



Estrada: *Lin. Alg. Appl.* 436 **2012** 4317-4328
Estrada: *Phys. Rev. E*, 85 **2012**, 066122

Communicability Paths





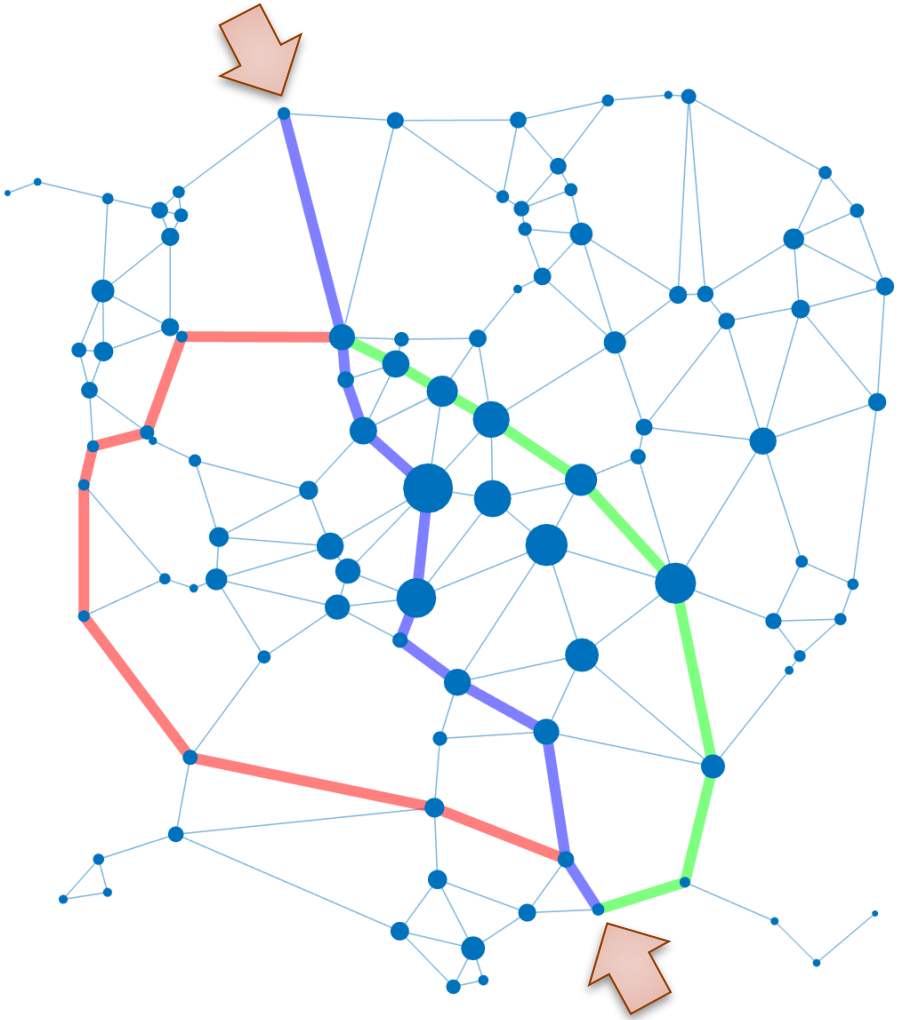
$$\sum_{\substack{(i,j)\in E \\ i,j\in P_1}} \left(\xi_{ij}\right)^{1/2} = 5.09$$

$$\sum_{\substack{(i,j)\in E \\ i,j\in P_2}} \left(\xi_{ij}\right)^{1/2} = 4.80$$

Remark: The shortest route between two nodes based on the communicability distance is the shortest path between those nodes that avoids the nodes with the highest 'cliquishness" in the graph.

Estrada: *Lin. Alg. Appl.* 436 **2012** 4317-4328
Estrada: *Phys. Rev. E*, 85 **2012**, 066122

Communicability Paths



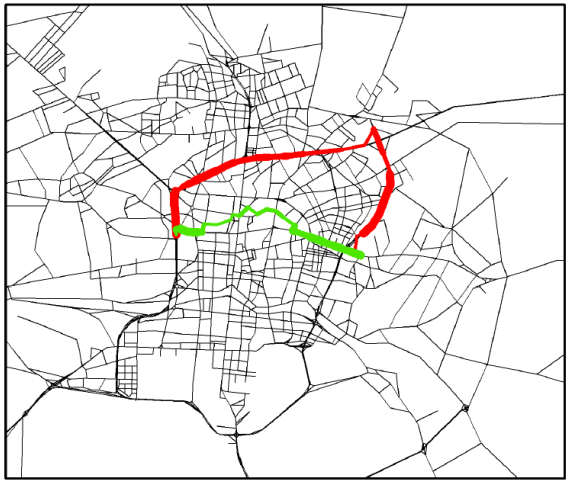
shortest path

shortest Euclidean distance

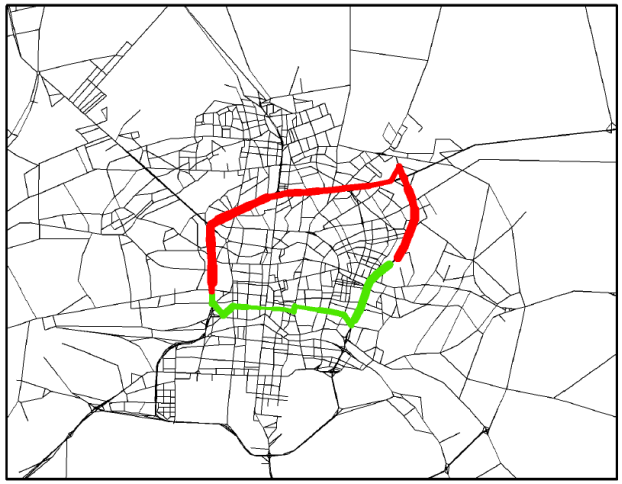
shortest communicability distance

Evidences?

Number of cars per hour in the morning at intersection points

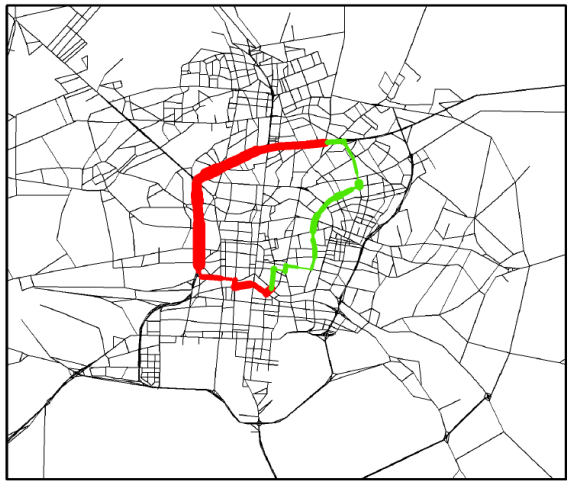


(A)

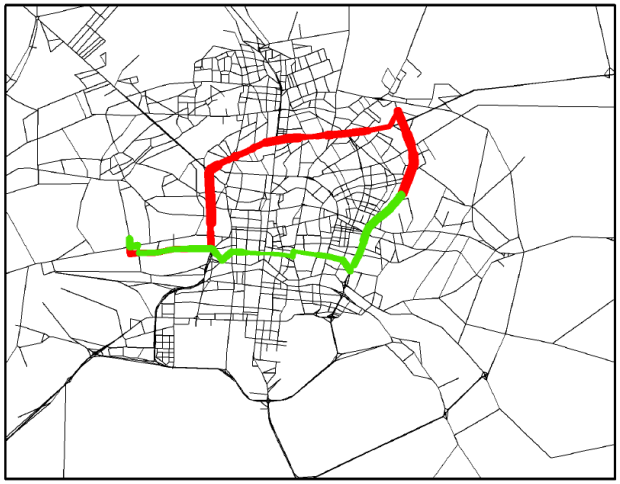


(B)

Isfahan, Iran
1,961,260 inhabitants



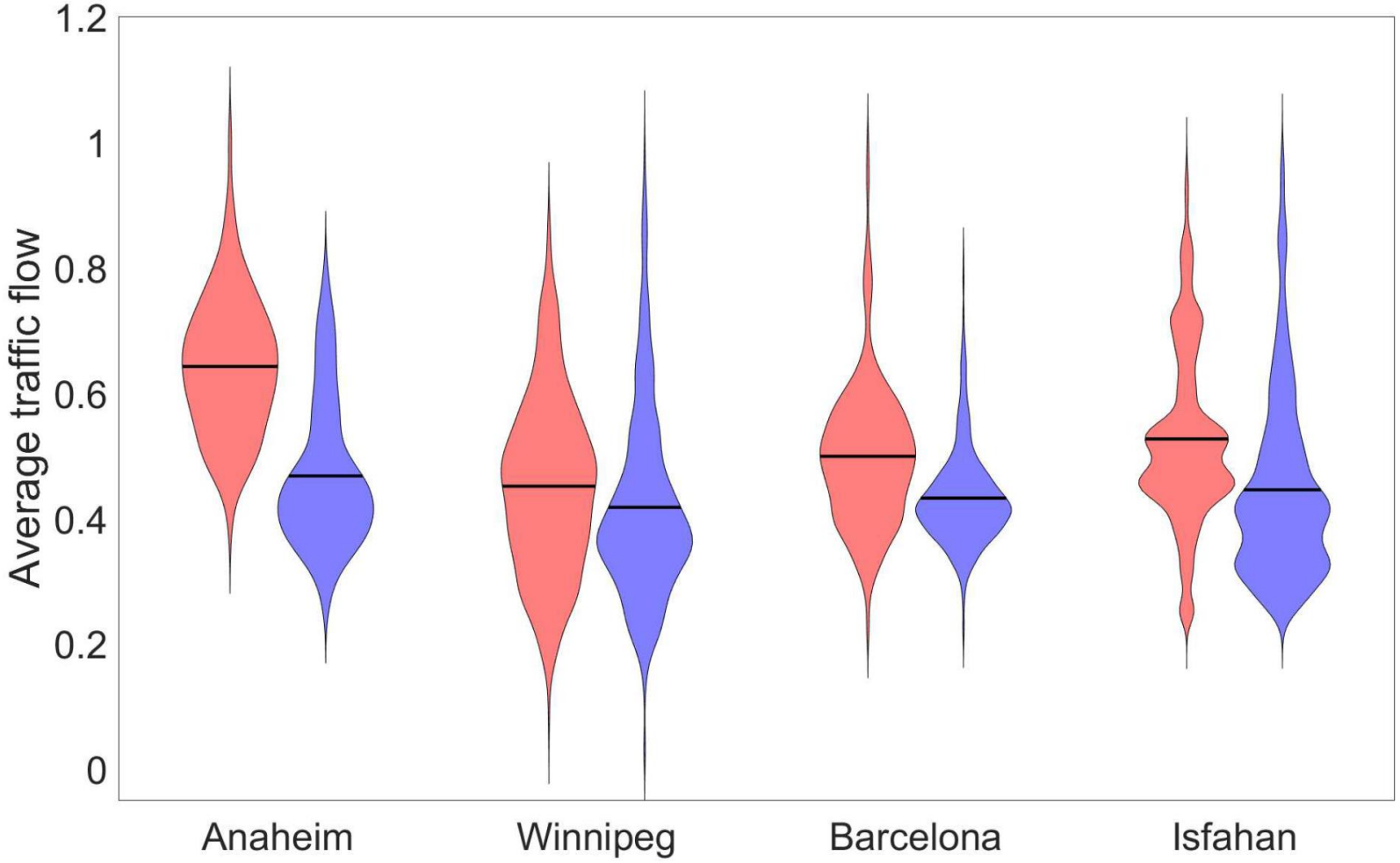
(C)



(D)

Estrada & Akbarzadeh: *submitted, 2017.*

Communicability geometry



Remark. *People do not use the shortest path but the shortest communicability paths!*

Estrada & Akbarzadeh: *submitted, 2017.*

Communicability angle

Definition:

$$\gamma_{pq} \overset{def}{=} \frac{G_{pq}}{\sqrt{G_{pp}G_{qq}}}$$

Theorem: The function γ_{pq} is the cosine of the Euclidean angle spanned by the position vectors of the nodes *p* and *q*.

$$\theta_{pq} = \cos^{-1}\frac{\vec{x}_p \cdot \vec{x}_q}{\|\vec{x}_p\|\|\vec{x}_q\|} = \cos^{-1}\frac{G_{pq}}{\sqrt{G_{pp}G_{qq}}}$$

Estrada & Hatano: *SIAM Rev*. 58, **2016,** 692-715 .

Communicability Geometry

Theorem: The communicability distance induces a natural embedding of a network into an n-dimensional Euclidean sphere of radius:

$$R^2 = \frac{1}{4}\left(c - \frac{(2-b)^2}{a} \right)$$

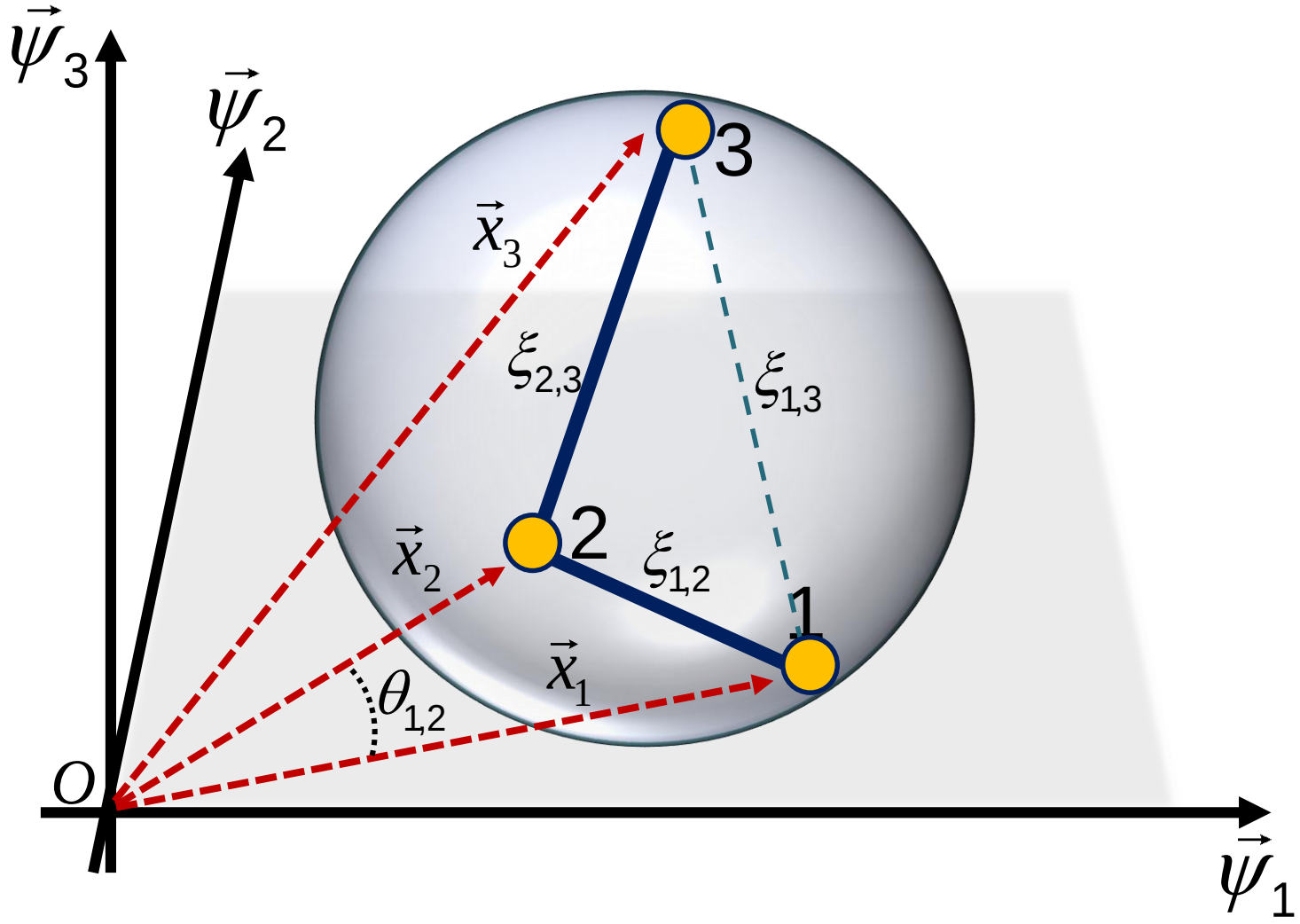
$$a = \vec{1}^T e^{-A} \vec{1} \qquad b = \vec{s}^T e^{-A} \vec{1} \qquad c = \vec{s}^T e^{-A} \vec{s}$$

Remark: The communicability distance matrix C is circum-Euclidean:

$$C = \vec{s}\vec{1}^T + \vec{1}\vec{s}^T - 2e^A \qquad \vec{s} = diag\left(e^A\right)$$

Estrada et al.: *Discrete Appl. Math.* **176** (2014) 53-77.
Estrada & Hatano: *SIAM Rev.* 58, **2016,** 692-715 .

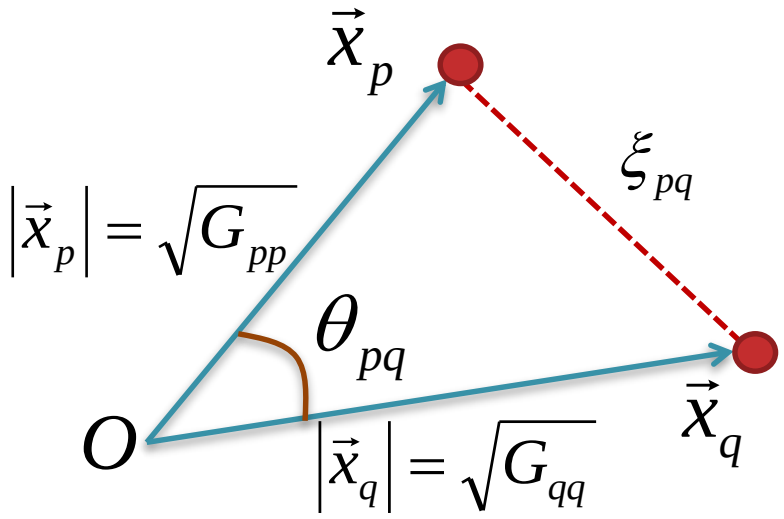
Communicability Geometry



Estrada & Hatano: *SIAM Rev*. 58, **2016,** 692-715 .

Communicability Geometry

Geometric interpretation



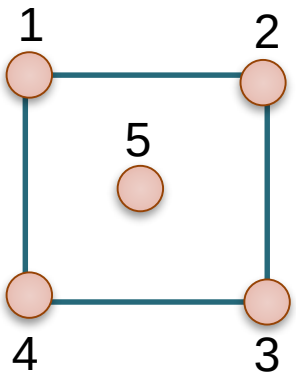
Communicability function

$$G_{pq} = \vec{x}_p \cdot \vec{x}_q$$

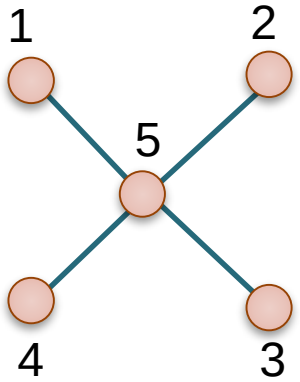
Estrada: *Lin. Alg. Appl*. **436** (2012) 4317-4328.
Estrada & Hatano: *SIAM Rev*. 58, **2016,** 692-715 .

Isospectrality: No problem!

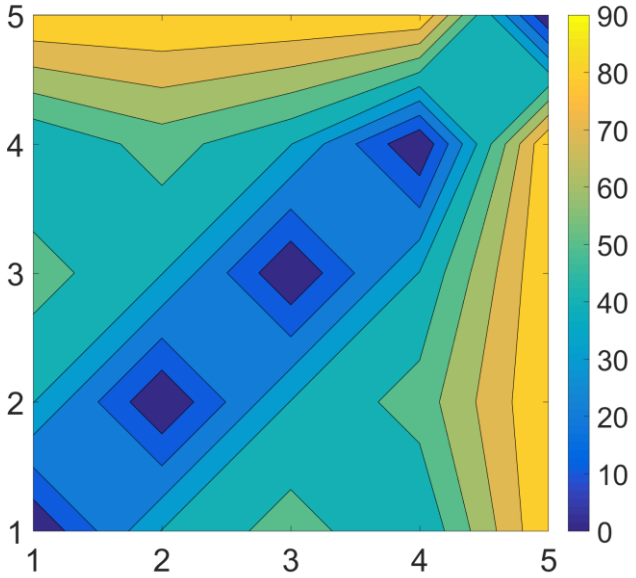
$G_1 =$



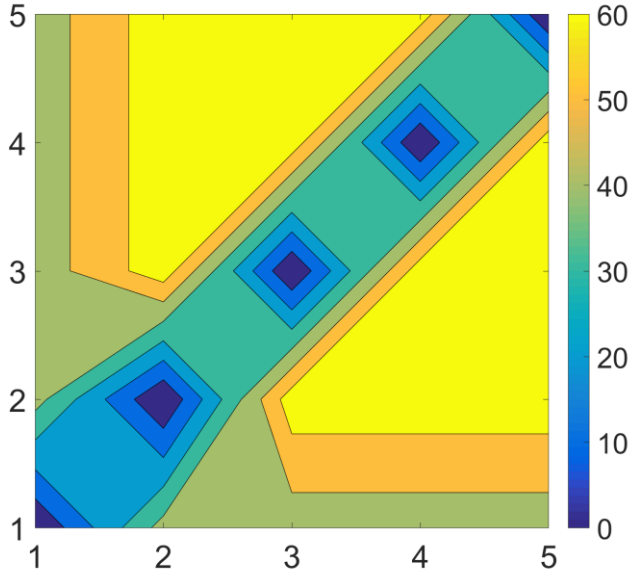
$G_2 =$



$\theta(G_1) =$



$\theta(G_2) =$



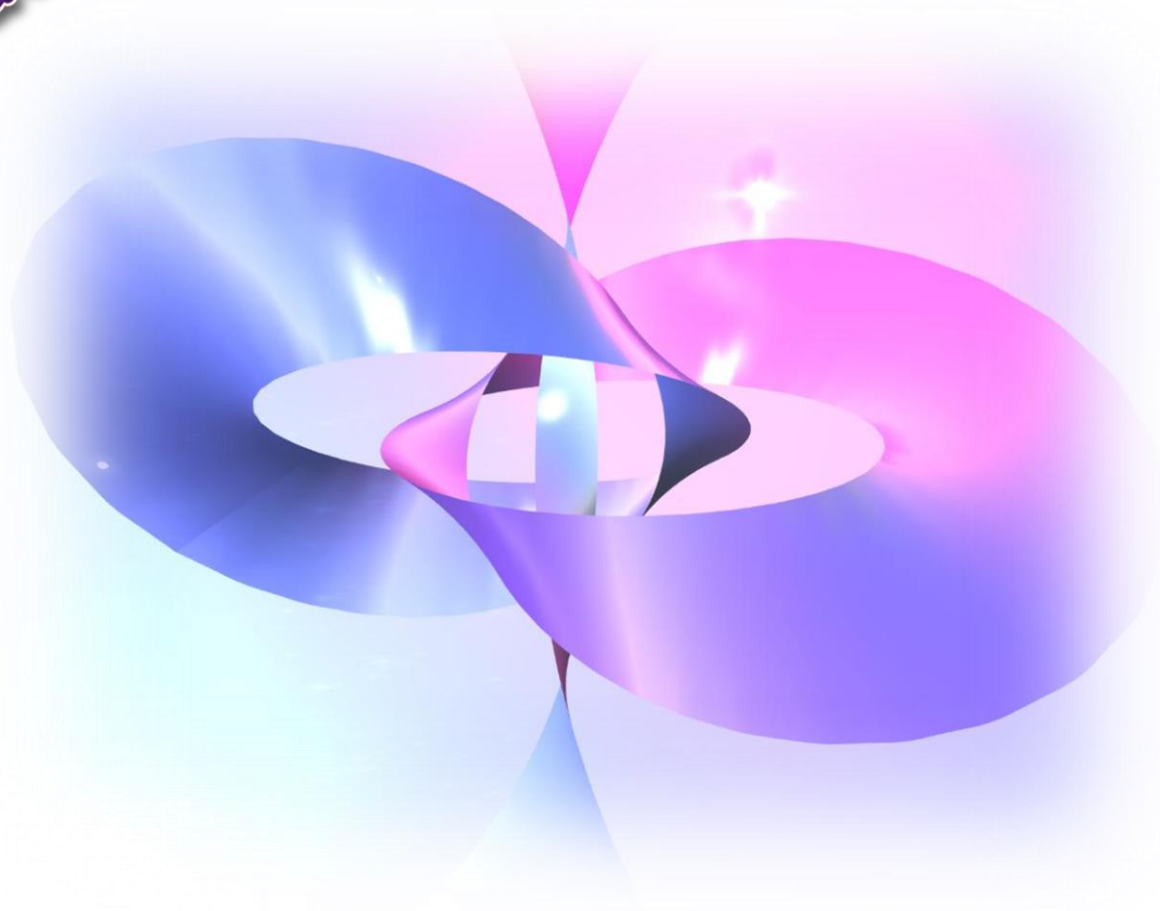
$\langle \theta(G_1) \rangle = 50.45$

$\langle \theta(G_2) \rangle = 45.71$

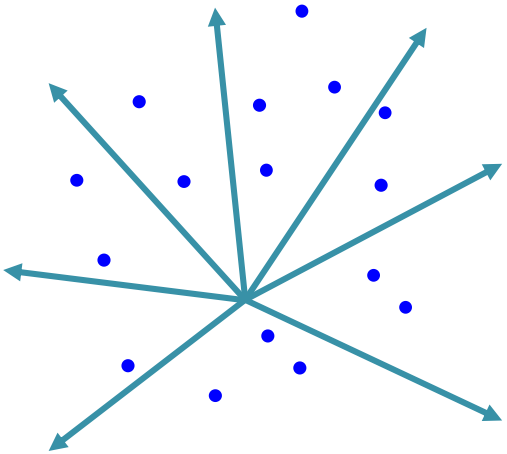
$R(G_1) = 1.0019$

$R(G_2) = 0.8802$

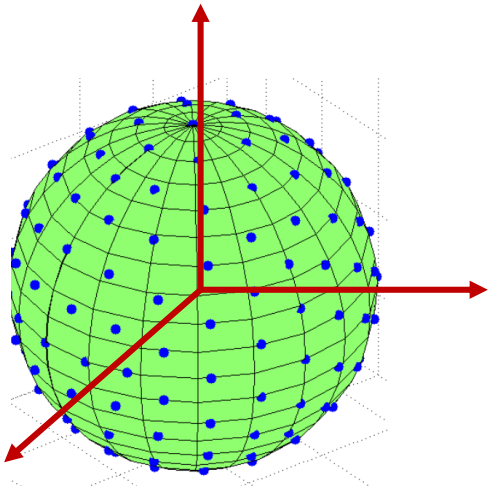
Dimensionality Reduction



Nonmetric Multidimensional Scaling (NMMDS)



Minimization of
Kruskal's *Stress* function



n-dimensional space

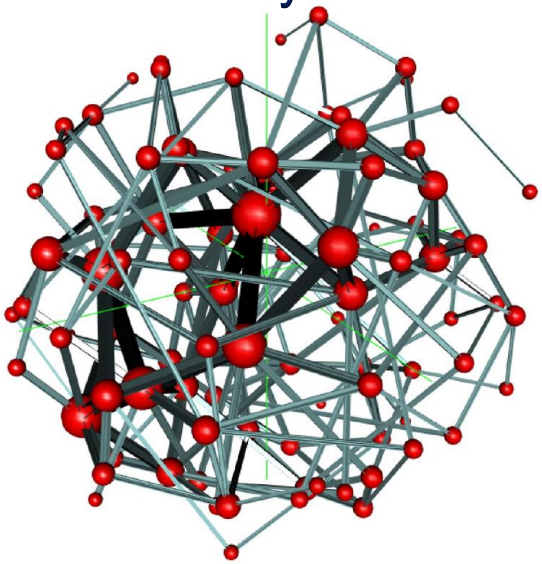
3-dimensional space

$$\theta = \left[\theta_{pq} \right]_{n \times n}$$

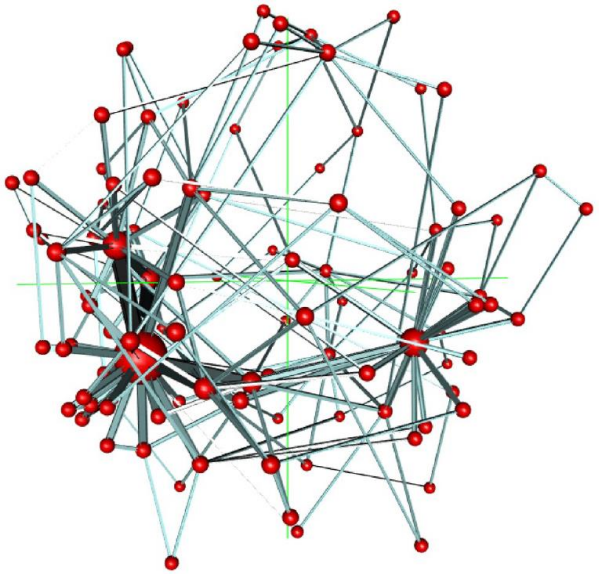
$$\tilde{\theta} = \left[\tilde{\theta}_{pq} \right]_{n \times n}$$

NMMDS

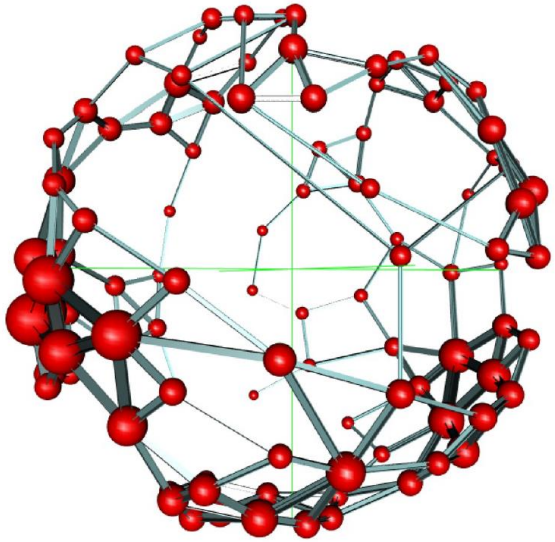
Erdős-Rényi network



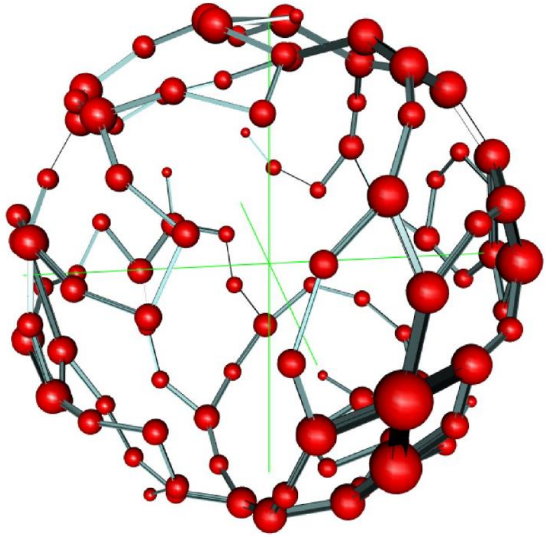
Barabási-Albert network



Gabriel graph



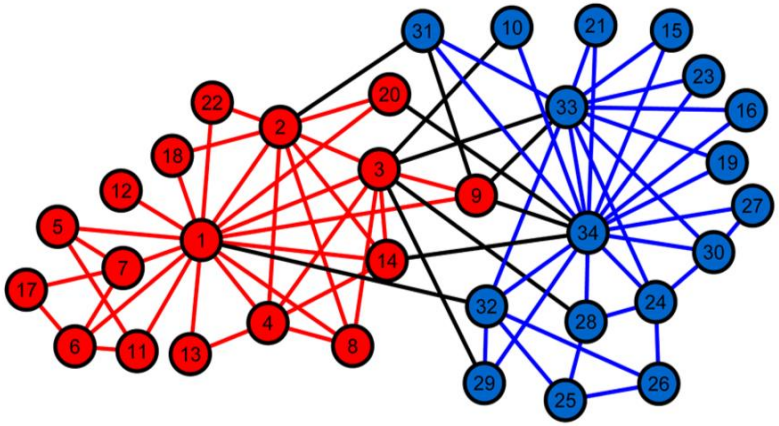
Relative neighbourhood graph

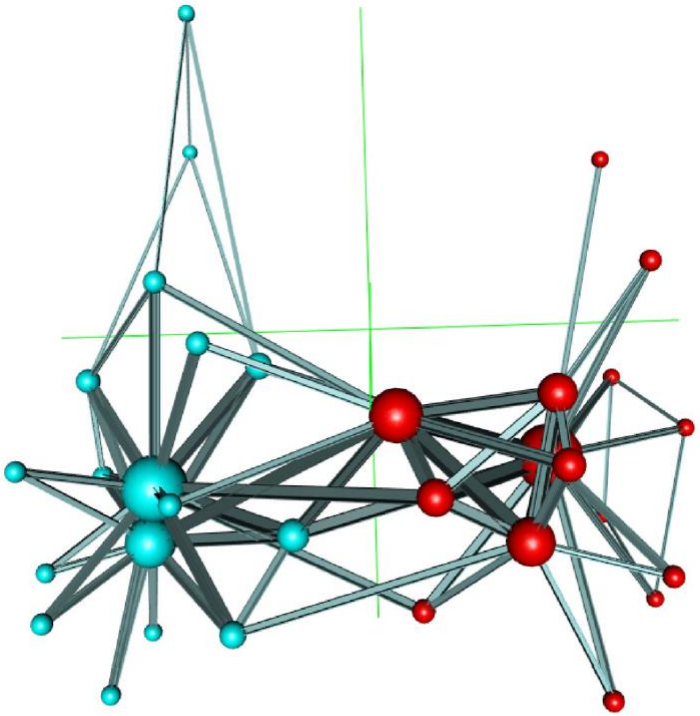


Pereda & Estrada: *arXiv:1804.05960*. **2018**.

NMMDS

Friendship network in a karate club

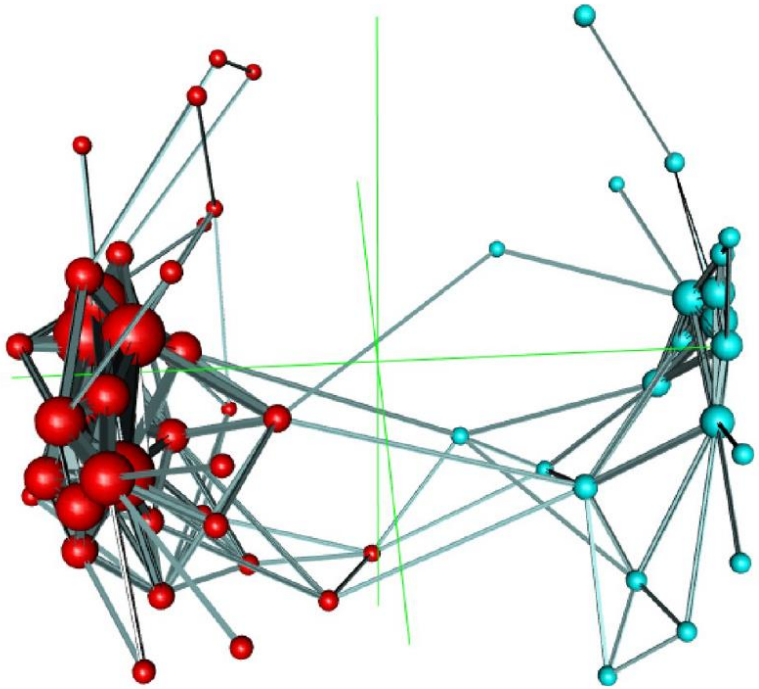


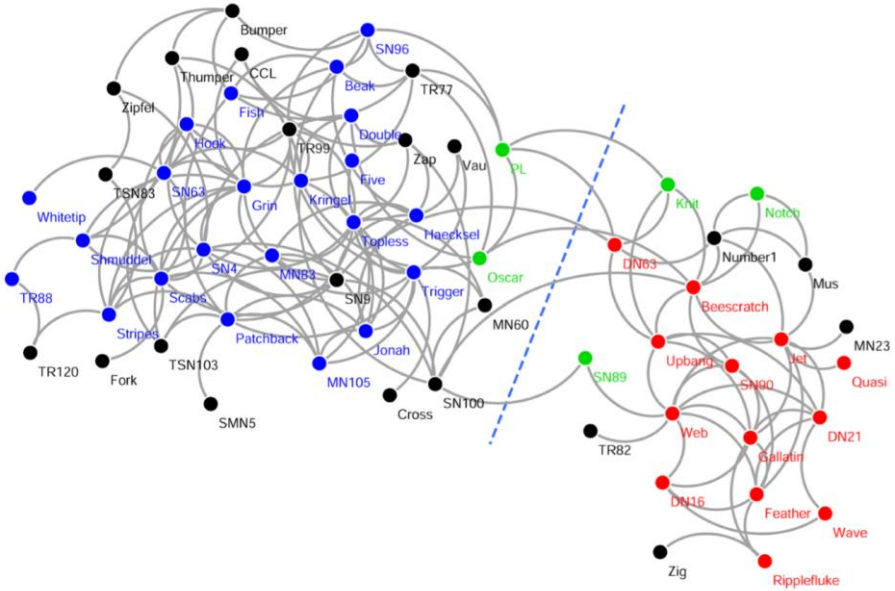


Pereda & Estrada: *arXiv:1804.05960*. **2018**.

NMMDS

Friendship network of bottlenose dolphins

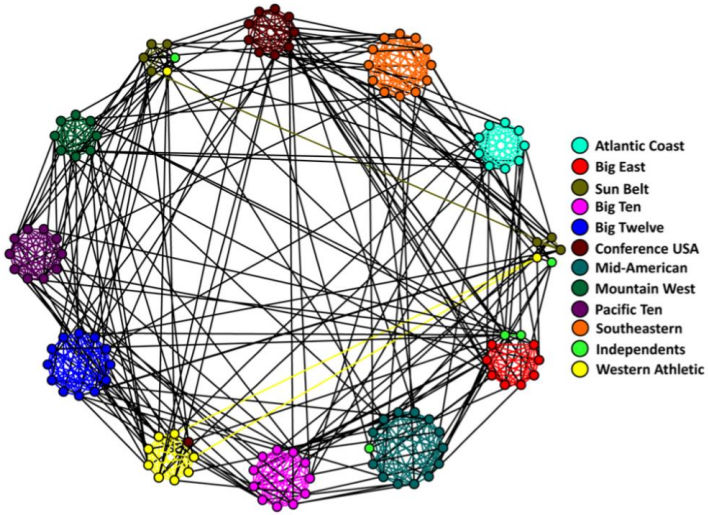


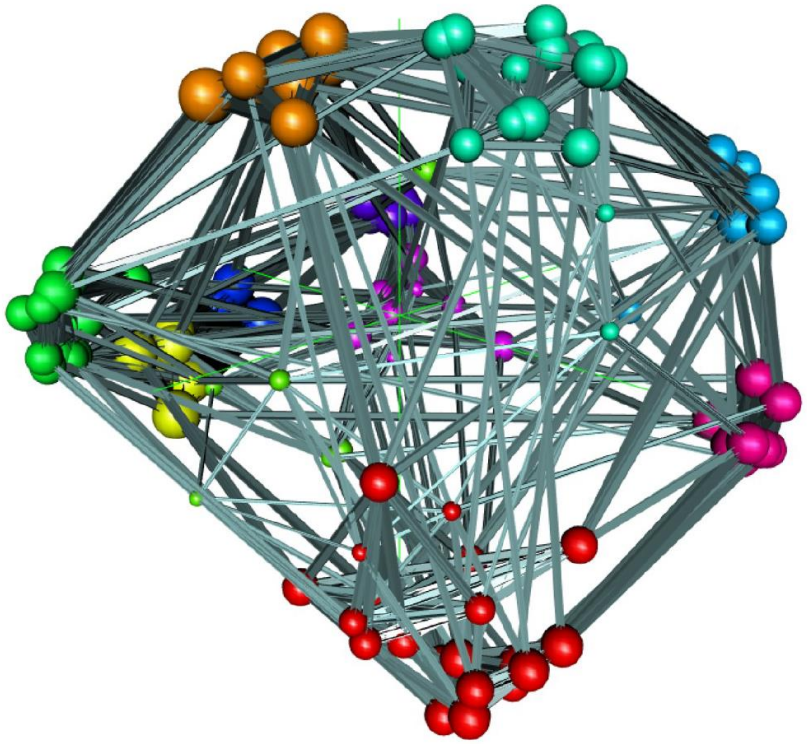


Pereda & Estrada: *arXiv:1804.05960*. **2018**.

NMMDS

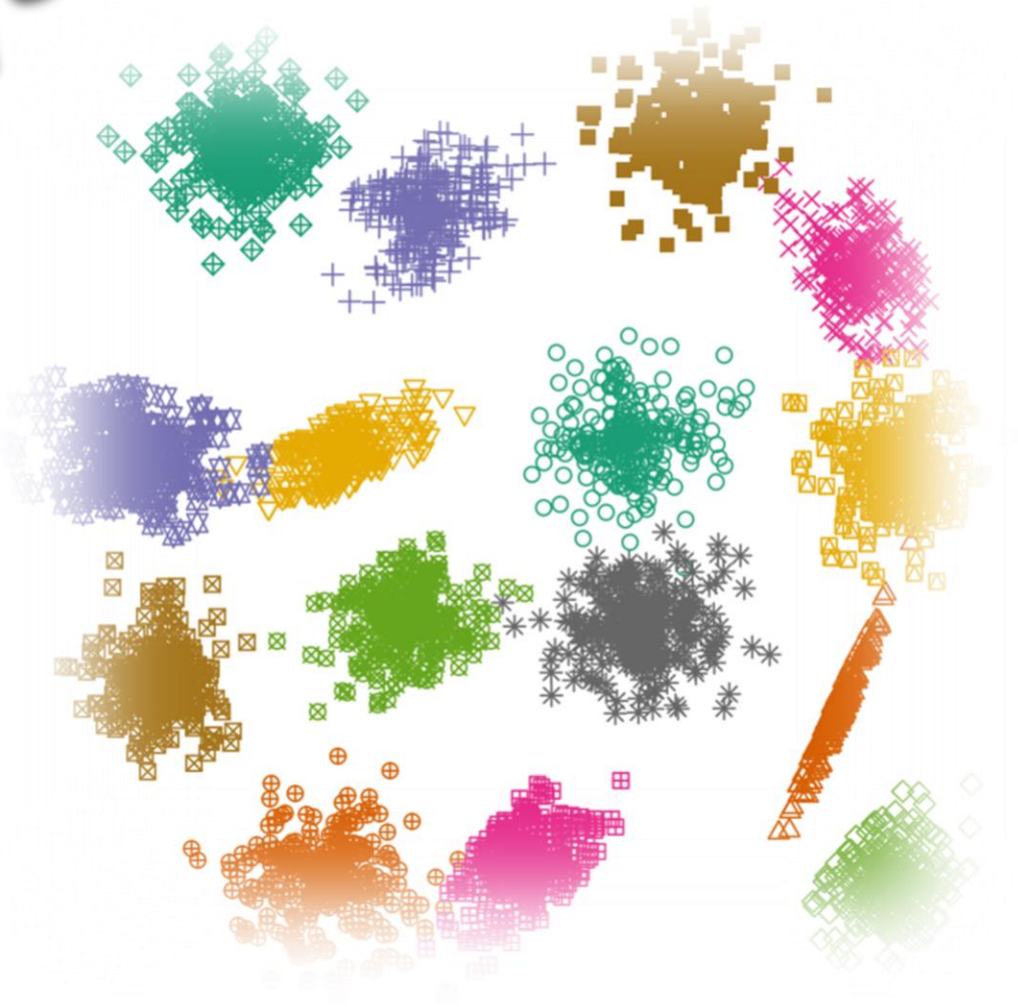
Network of American football games between Division IA colleges





Pereda & Estrada: *arXiv:1804.05960*. **2018**.

Clustering Analysis



Networks with ground truth

	ground truth		Silhouette			Calinski-Harabasz			Davies-Bouldin			Methods	
network	C	Q	C	NMI	Q	C	NMI	Q	C	NMI	Q	NMI	Q
Karate	2	0.37	**2**	**1.00**	**0.37**	**2**	**1.00**	**0.37**	**2**	**1.00**	**0.37**	1.00	0.37
Dolphins	2	0.38	**2**	**0.89**	**0.38**	**2**	**0.89**	**0.38**	**2**	**0.89**	**0.38**	0.89	0.38
Football	12	0.55	11	0.91	0.66	12	0.91	0.64	13	0.90	0.66	0.91	0.66
PolBooks	2	0.41	**2**	**0.61**	**0.44**	7	0.58	0.45	**2**	**0.61**	**0.44**	0.60	0.44
PolBlogs	2	0.41	**2**	**0.72**	**0.52**	-	-	-	**2**	**0.72**	**0.52**	0.71	0.52
Average				0.83						0.82			

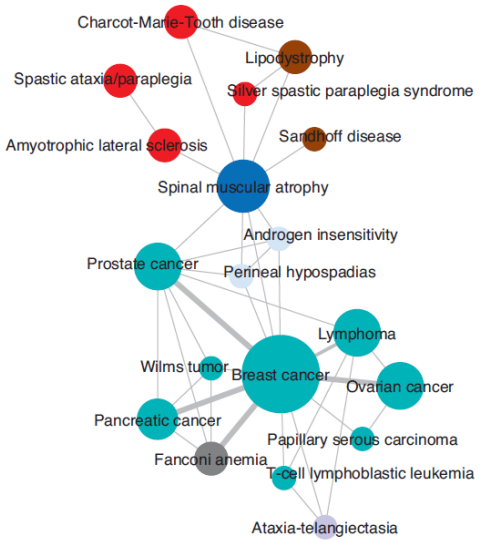
	Louvain		FastGreedy		Infomap		Eigenvector		LP		NMI
network	NMI	Q	NMI	Q	NMI	Q	NMI	Q	NMI	Q	
Karate	0.59	0.42	0.69	0.38	0.70	0.40	0.68	0.39	0.70	0.40	0.73
Dolphins	0.48	0.52	0.61	0.50	0.50	0.52	0.54	0.49	0.69	0.50	0.64
Football	0.88	0.60	0.70	0.55	0.92	0.60	0.70	0.49	0.92	0.60	0.83
PolBooks	0.51	0.52	0.53	0.50	0.49	0.52	0.52	0.47	0.57	0.50	0.54
PolBlogs	0.63	0.43	0.65	0.43	0.48	0.42	0.69	0.42	0.69	0.43	0.64
Average	0.62		0.64		0.62		0.63		0.71		

C: # clusters. NMI: Normalized Mutual Information. Q: Newman modularity

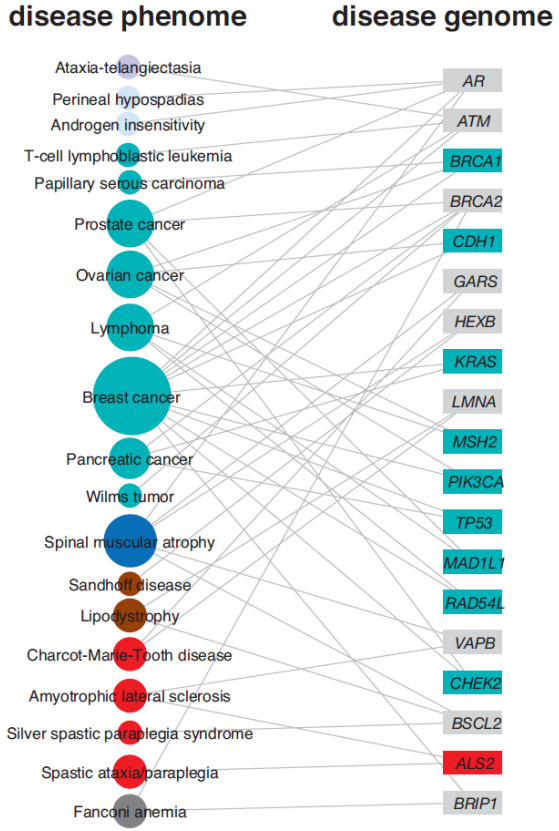
Networks without ground truth

Human genetic diseases

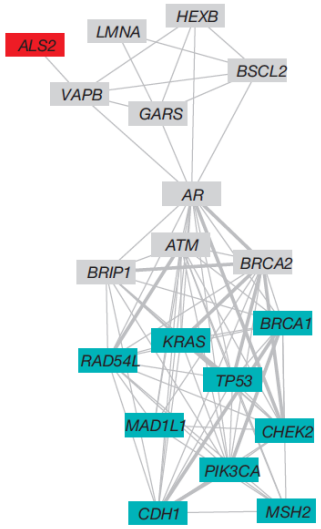
Human Disease Network (HDN)



DISEASOME



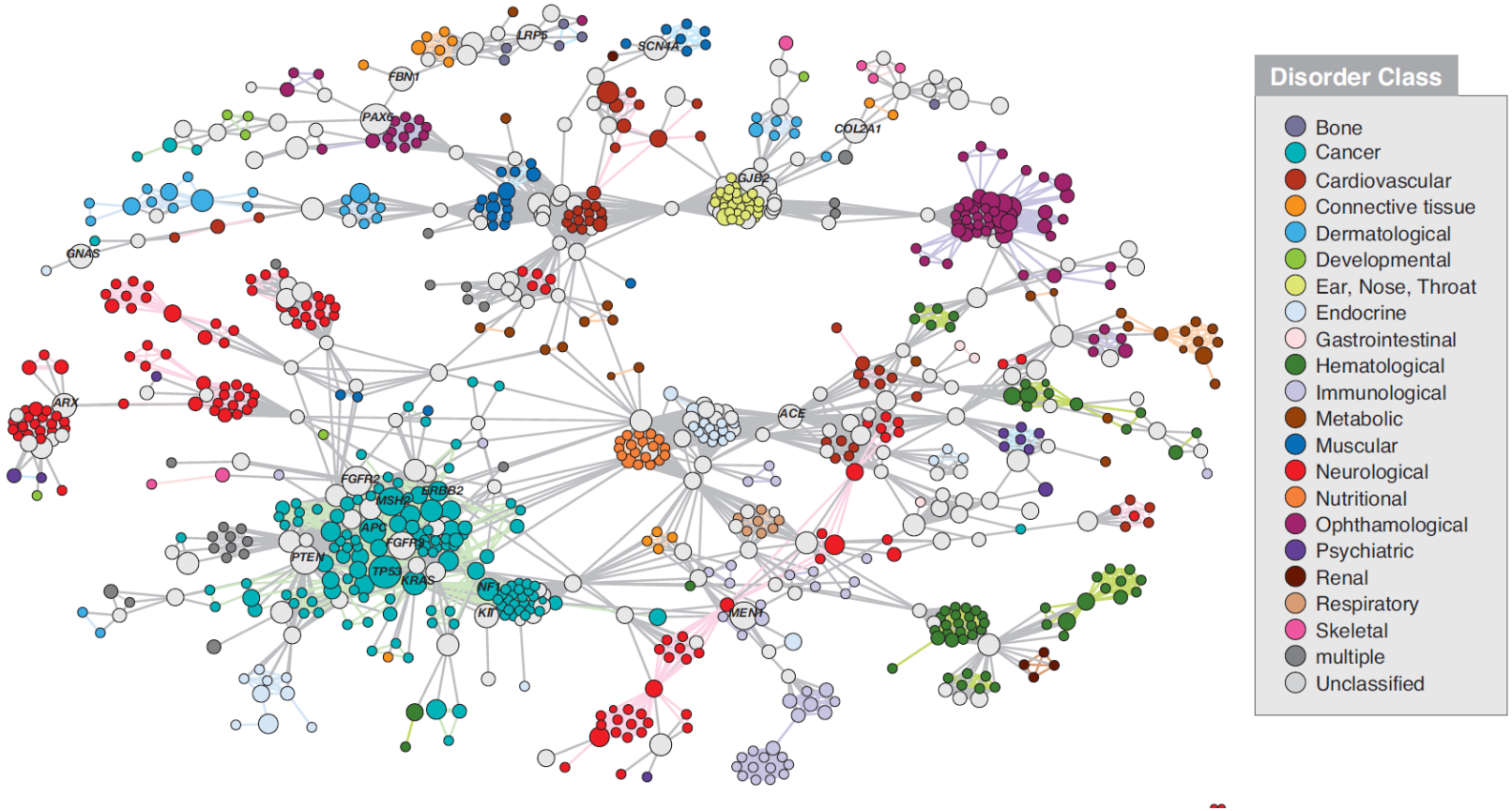
Disease Gene Network (DGN)



Goh KI, et al. *Proc. Nat. Acad. Sci.* 104, **2007**, 8685-90.

K-Means Clustering Analysis

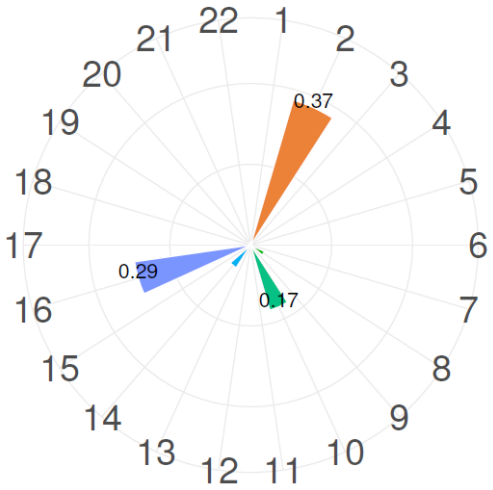
Network of gene co-participation in genetic diseases



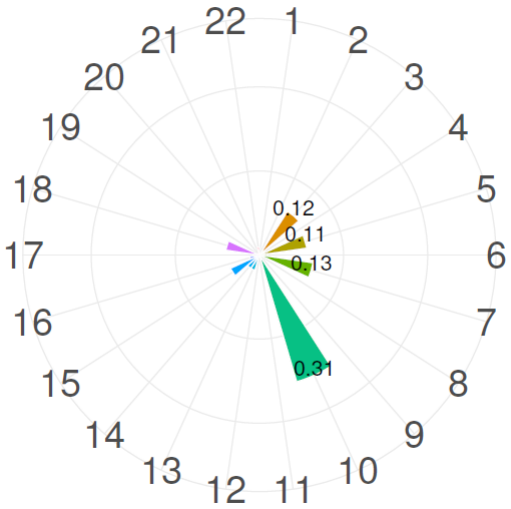
Goh KI, et al. *Proc. Nat. Acad. Sci*. 104, **2007**, 8685-90.

Clusters found

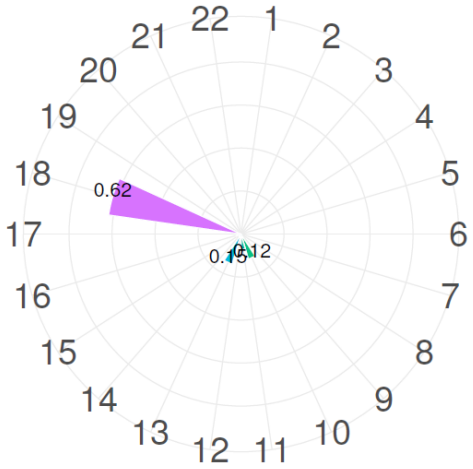
Cluster # 1



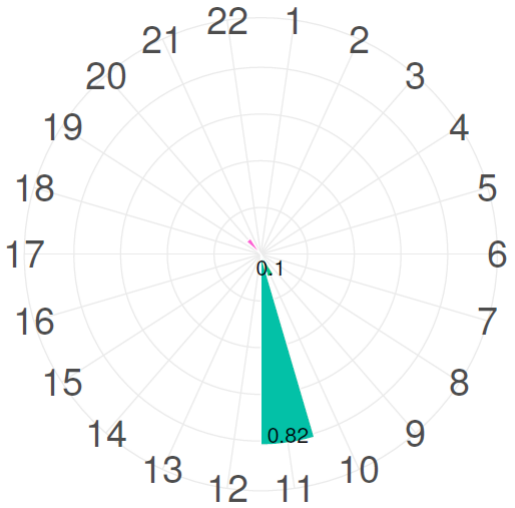
Cluster # 2



Cluster # 3



Cluster # 4

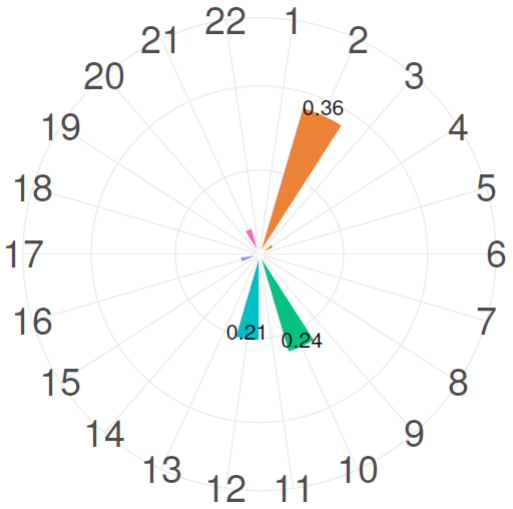


1: Bone.
2: Cancer.
3: Cardiovascular.
4: Connective tissue.
5: Dermatological.
6: Developmental.
7: Ear/Nose/Throat.
8: Endocrine.
9: Gastrointestinal.
10: Mixed.
11: Hematological.
12: Immunological.
13: Metabolic.
14: Multiple body parts.
15: Muscular.
16: Neurological.
17: Nutritional.
18: Ophtalmological.
19. Psychiatrical.
20. Renal.
21: Respiratory.
22: Skeletal.

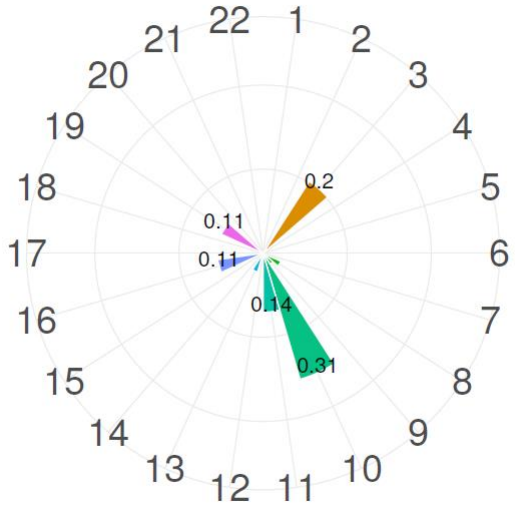
Pereda & Estrada: *arXiv:1804.05960*. **2018**.

Clusters found

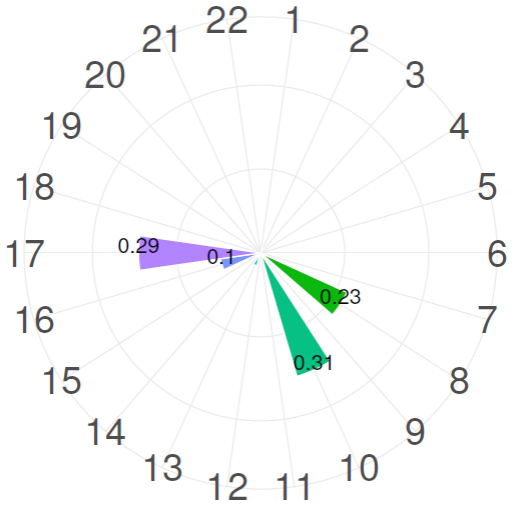
Cluster # 5



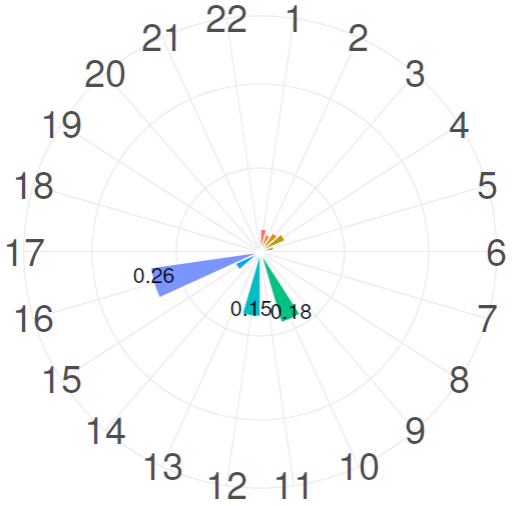
Cluster # 6



Cluster # 7



Cluster # 8

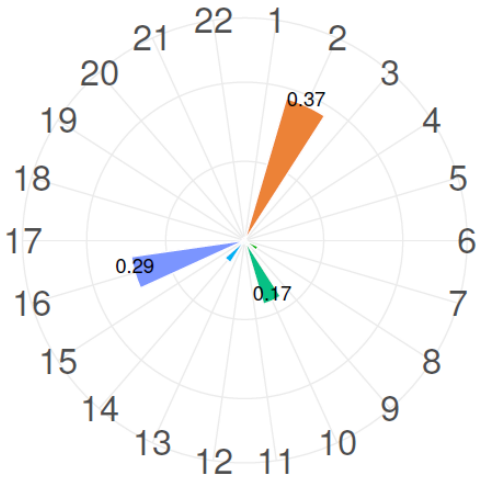


1: Bone.
2: Cancer.
3: Cardiovascular.
4: Connective tissue.
5: Dermatological.
6: Developmental.
7: Ear/Nose/Throat.
8: Endocrine.
9: Gastrointestinal.
10: Mixed.
11: Hematological.
12: Immunological.
13: Metabolic.
14: Multiple body parts.
15: Muscular.
16: Neurological.
17: Nutritional.
18: Ophtalmological.
19. Psychiatrical.
20. Renal.
21: Respiratory.
22: Skeletal.

Pereda & Estrada: *arXiv:1804.05960*. **2018**.

Analysis of cluster # 1

Cluster # 1



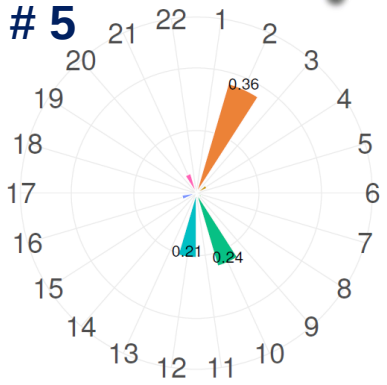
Hypothesis: Genes involved in neurological disorders which are in cluster 1 can also be involved in CANCER.

No.	gene	neurological
1	NDRG1	Charcot-Marie-Tooth
2	FGF14	Spinocereballar ataxia
3	NEFH	ALS
4	PPP2R2B	Spinocereballar ataxia
5	SLC25A22	epilepsy
6	GABRA1	epilepsy
7	JPH3	Huntington's
8	GJB1	Charcot-Marie-Tooth
9	UCHL1	Parkinson
10	DNM2	Charcot-Marie-Tooth
11	TDP1	Spinocereballar ataxia
12	SOD1	ALS
13	PARK7	Parkinson
14	LRRK2	Parkinson
15	KIF1B	Charcot-Marie-Tooth
16	HSPD1	Spastic ataxia/paraplegia
17	NR4A2	Parkinson
18	Rab7	Charcot-Marie-Tooth
19	SNCAIP	Parkinson

Pereda & Estrada: *arXiv:1804.05960*. **2018**.

Analysis of cluster # 5

Cluster # 5



Hypothesis: Genes involved in "mixed" disorders which are in cluster 5 can also be involved in CANCER.

No.	gene	"grey" diseases
1	ABCA1	Cerebral amyloid angiopathy, Coronary artery disease, HDL cholesterol level QTL, Tangier disease
2	ESR1	Estrogen_resistance, HDL cholesterol level QTL, Migraine
3	ALOX5	Asthma, Atherosclerosis
4	IL10	Graft -versus-host disease, HIV, Rheumatoid arthritis
5	IL13	Allergic rhinitis, Asthma
6	CIITA	Bare lymphocyte syndrome, Multiple_sclerosis, Rheumatoid arthritis
7	PTPRC	Multiple sclerosis, Severe combined immunodeficiency
8	BDNF	Central hypoventilation_ syndrome, Memory impairment, Obsessive-compulsive disorder
9	PLA2G7	Asthma, Atopy, Platelet defect/deficiency
10	CD36	Malaria, Platelet defect/deficiency

Pereda & Estrada: *arXiv:1804.05960*. **2018**.

Conclusions

- Networks are naturally embedded into *n*-dimensional spheres
- The embedding is induced by the Euclidean geometry emerging from the communicability function
- The communicability distance and communicability angle are similarity measures for pairs of vertices in networks
- The hyperspherical embedding of networks reveal their community structure, which can be extracted using standard machine learning techniques.

Works for collaboration

- ~~Extension to multiplexes (done)~~
- Extensions to hypergraphs
- Extensions to simplicial complexes
- Hyperspherical embedding and graph planarity

- Applications of other machine learning techniques
- Applications to other real-world problems

Some References

Communicability functions:

- Estrada & Rodríguez-Velázquez: *Phys. Rev. E* 71, **2005**, 056103
- Estrada & Hatano: *Phys. Rev. E* 77, **2008**, 036111
- Estrada, Hatano & Benzi, *Phys. Rep*. 514 **2012**, 89-119
- Estrada & Higham, *SIAM Rev*. 52, **2010**, 696-714
- Estrada: *J. Theor. Biol*. 263 **2010**, 556-565.

Communicability geometry:

- Estrada: *Lin. Alg. Appl*. 436 **2012** 4317-4328
- Estrada: *Phys. Rev. E*, 85 **2012**, 066122
- E. Estrada, Sanchez-Lirola, de la Peña, *Discr. Appl. Math*. 176 **2014**, 53-77.
- Estrada & Arrigo, *SIAM J. Appl. Math*. 75 **2015**, 1725-1744.
- Estrada & Hatano: *SIAM Rev. 58,* **2016** 58, 692-715.
- Estrada, Vargas-Estrada & Ando: *Phys. Rev. E 92* **2015** 052809.

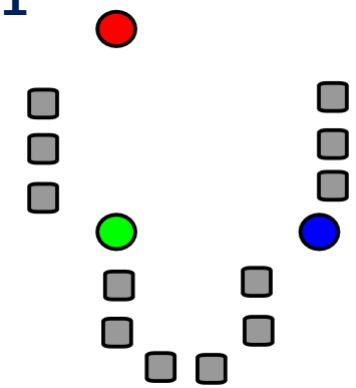
All refs. can be downloaded from:

www.estradalab.org

Thank you!

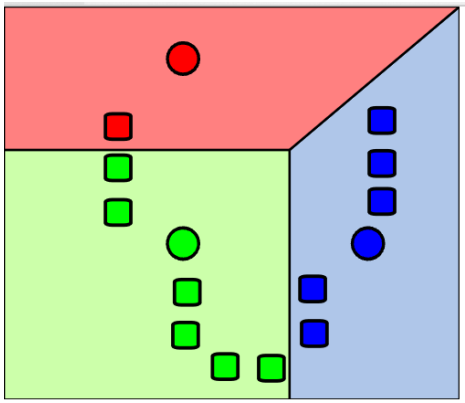
K-Means Clustering Analysis

Step 1



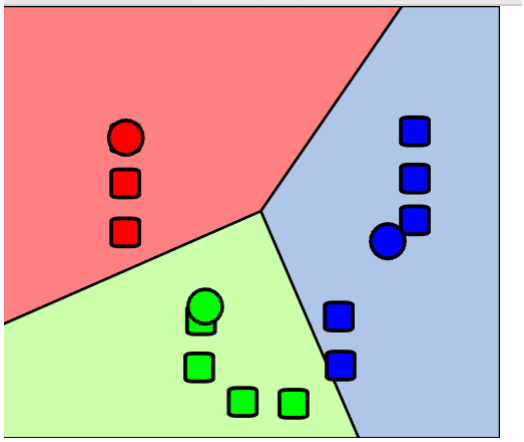
randomly generate
k initial “means”

Step 2



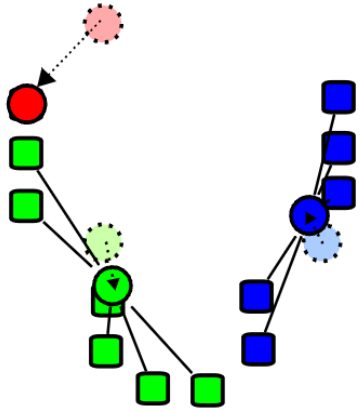
k clusters are created by associating every observation with its nearest mean

Step 3



create new means by using the centroid of each cluster

Step 4



repeat steps 2 and 3 until convergence is reached

Clustering Validation Indices (CVI)

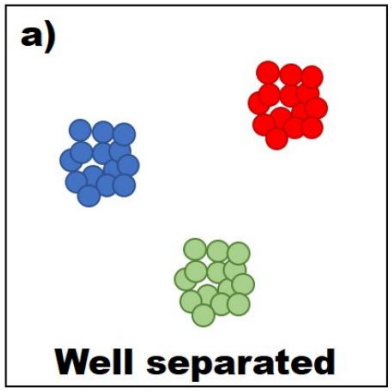
Ingredients of CVIs

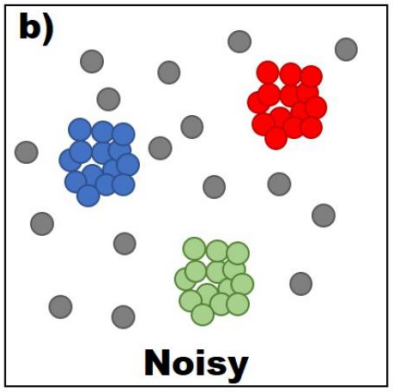
Indices

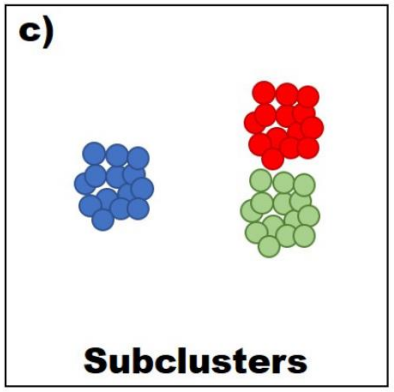
- **Calinski-Harabasz**
- **Silhouette**
- **Davies-Bouldin**

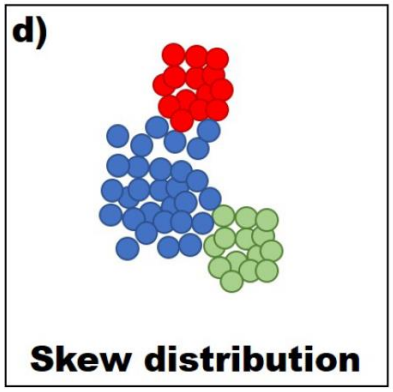
Liu Y, *et al*. In *2010 IEEE 10th International Conference on Data Mining (ICDM)*, **2010** Dec 13 (pp. 911-916). IEEE.

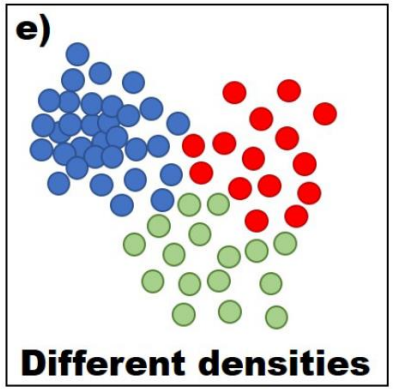
Cluster Validity Indices

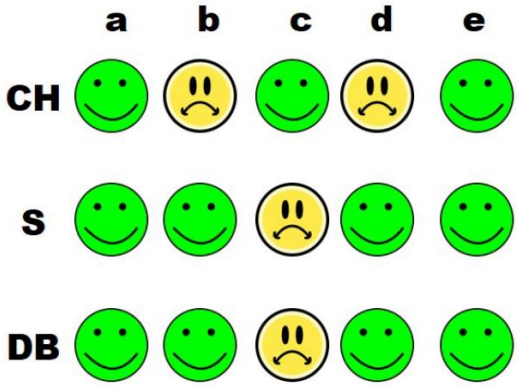










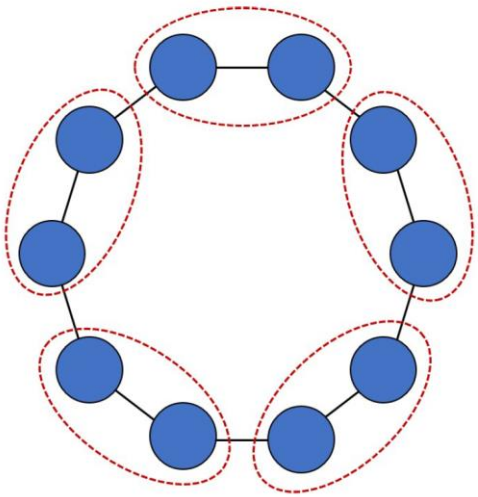


CH: Calinski-Harabasz. S: Silhouette. DB: Davies-Bouldin

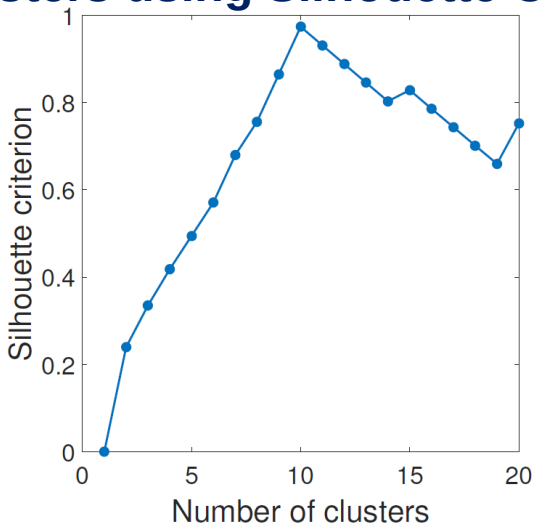
Liu Y, *et al*. In *2010 IEEE 10th International Conference on Data Mining (ICDM)*, **2010** Dec 13 (pp. 911-916). IEEE.

Resolution limit

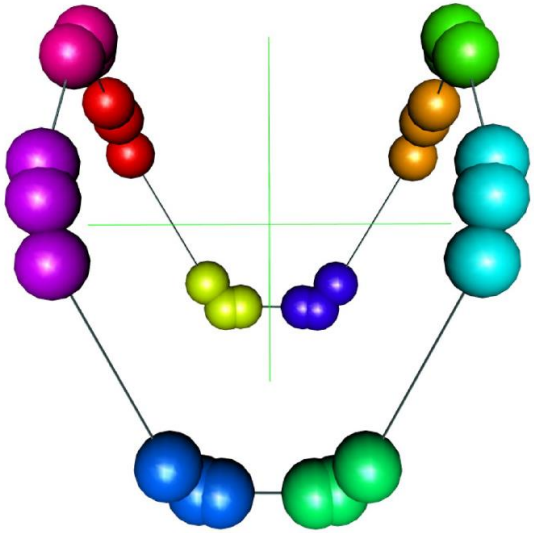
Caveman network



Clusters using Silhouette CVI



"Best" Silhouette clustering



Pereda & Estrada: *arXiv:1804.05960*. **2018**.