Community Detection in University Email Networks.

David Manuel de Sousa Rodrigues

Thesis supervisors
Professor Jorge Louçã, Lisbon University Institute
Professor John Symons, University of Texas in El Paso

Master Program in Complexity Sciences
June 2009

Summary

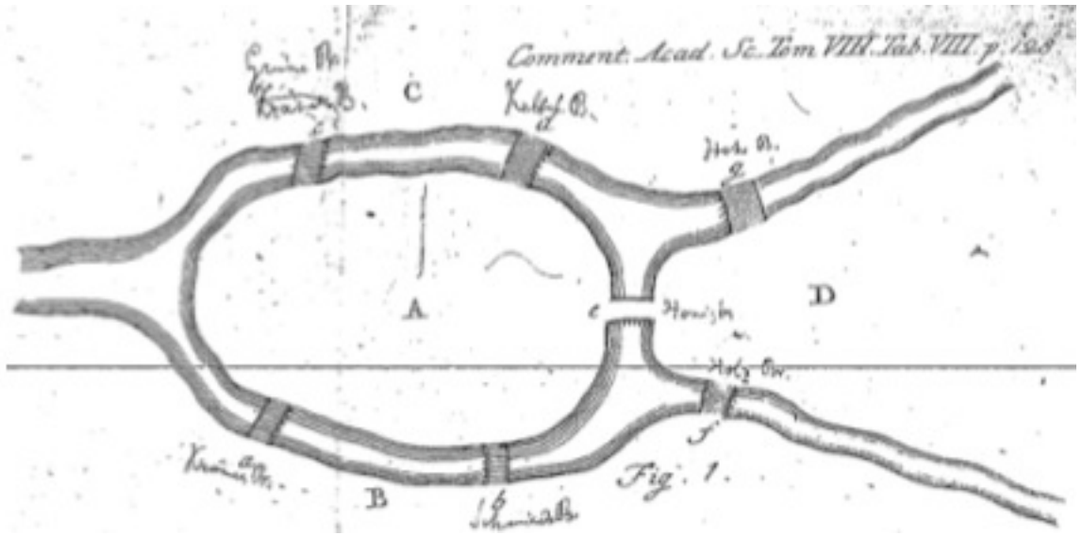
- Objectives
- State of the Art
- The ISCTE Email Case Study
  - Community Detection in Email Networks.
  - Information Variation to assess different clusterings.
- Conclusions

Objectives

- Identify the available mechanisms and methods to characterize social networks

- Understand the latent informal communication structure in email networks:
  - Case Study: ISCTE email network
  - Detect communities that regularly use the email network
  - Identify the macro properties of the network
  - Model the communication process.

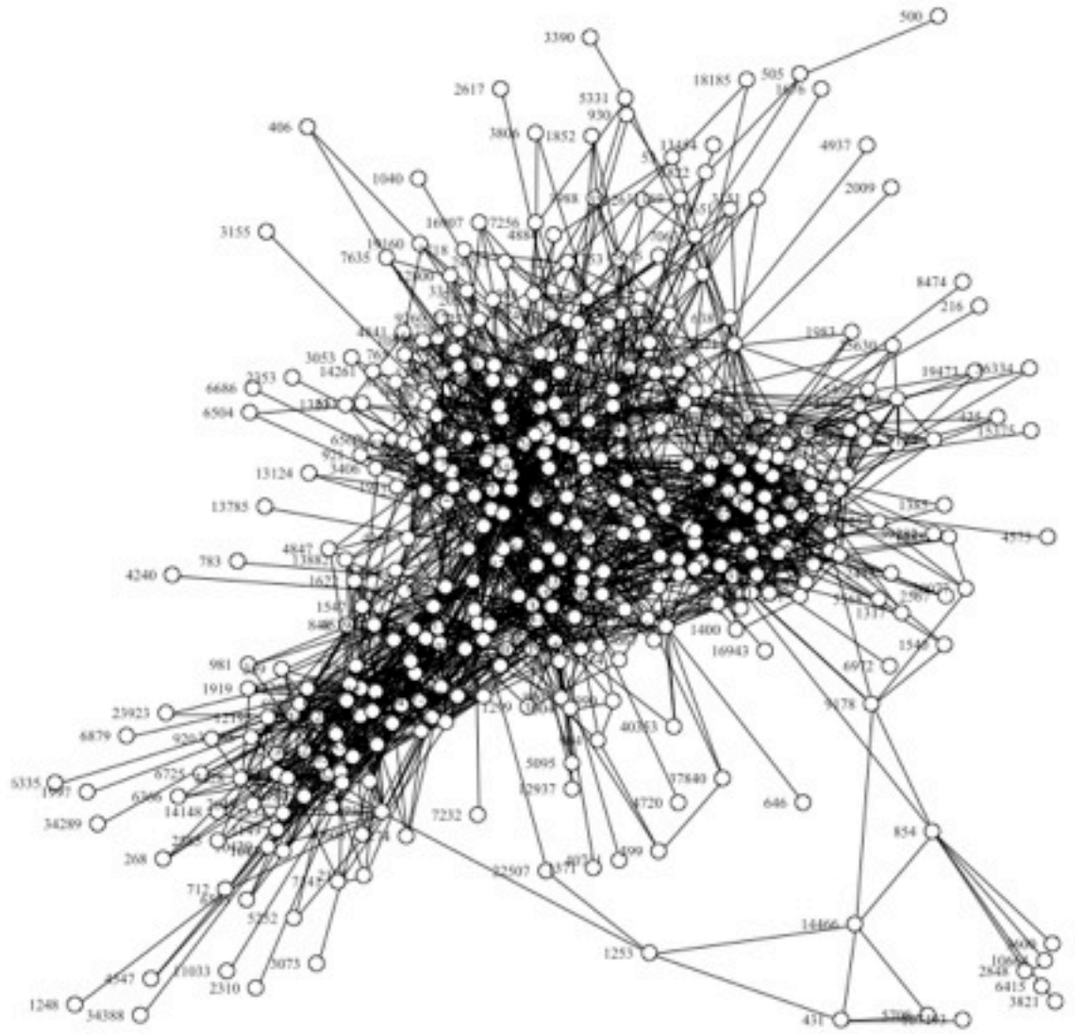
State of the Art



(Euler, 1741)

Community Detection Algorithms.

- Hierarchical Algorithms (Global):
  - Girvan-Newman
  - Clauset-Newman-Moore
- k-core analysis (Local).
- Clique Percolation (Local).



Girvan-Newman Algorithm (Girvan, 2001).

1. Calculate the edge betweenness for all edges in the graph.

2. Remove the edge with highest value of edge betweenness.

3. Recalculate edge betweenness for all edges affected by the previous removal.

4. Repeat from 2 until there are no edges left in the graph.

Clauset-Newman-Moore Algorithm (Clauset, 2004)

1. Calculate the increase in modularity for every possible join in the network

2. Select the join that maximizes the increase in modularity and merge both communities

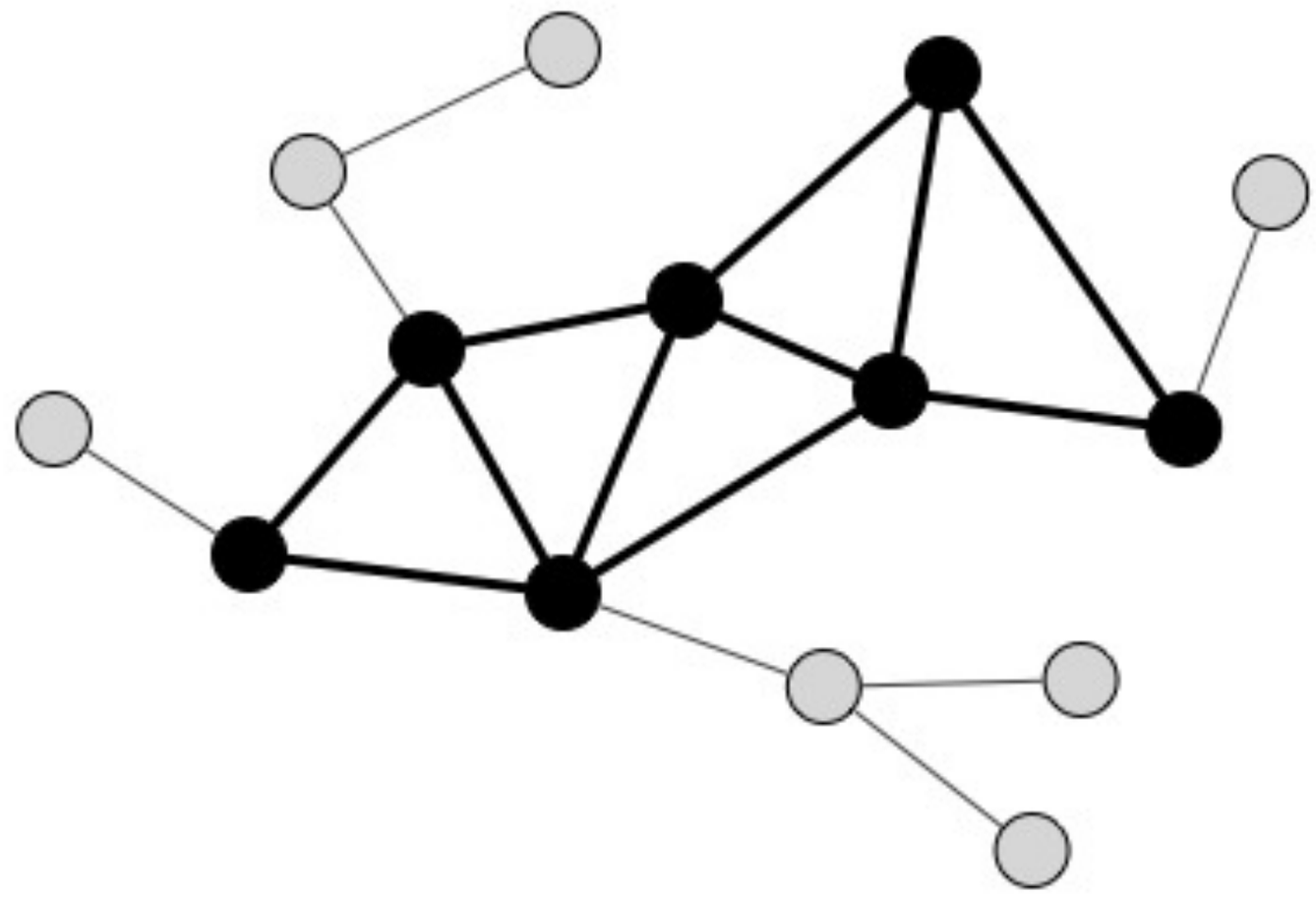
3. Repeat until there's only one community.

k-core analysis

- A *k*-core is the largest subgragh where every node has at least *k* connection
  - Iteratively remove all nodes with degree less than *k*
  - It results in an hierarchical structure of *k*-cores where each is encapsulated inside the next similarly to a russian doll (Dorogovtsev, 2005).

Clique Percolation (Palla, 2005).

- Allows for the overlapping of communities
- a community is constituted by the chain of adjacent *k*-cliques (sharing k-1 nodes)
- Two cliques are connected if they belong to the same *k*-clique chain.



Variation of Information to assess clustering distance (Meilă, 2007).

- Uses Association Matrixes

- For 2 clusterings C and C' the association matrix is a *k x k'* that the *kk'* elements gives the number of elements of Ck that are in C'k'

$$n_{kk'} = |C_k \cap C'_{k'}|$$

0	0	1	3	0	0	0	0	0	0	0	0	0	0
0	0	13	2	1	54	0	11	0	0	28	5	0	0
0	0	9	8	19	0	0	0	1	0	0	0	0	0
0	0	1	1	0	5	0	0	0	83	0	0	0	0
5	0	10	17	0	1	0	0	0	0	0	0	0	0
2	0	4	0	0	0	0	0	43	0	0	1	0	0
0	0	1	2	0	0	0	0	0	0	0	0	0	0
0	2	1	0	0	0	17	0	0	0	0	0	24	0
0	6	0	0	0	0	0	0	0	0	0	0	0	0
0	0	2	2	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	2
0	6	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0

Variation of Information to assess clustering distance (Meilă, 2007).

- Probability of node k belonging to Cluster C_k

$$P(k) = \frac{n_k}{n}$$

- Entropy of this probability

$$H(C) = -\sum_{k=1}^{K} P(k) \log(P(k))$$

- Mutual Information of 2 Clusterings *C* and *C'*

$$P(k,k') = \frac{|C_k \cap C'_{k'}|}{n}$$

$$I(C,C') = \sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k,k') \log \frac{P(k,k')}{P(k)P'(k')}$$

- Meila defined the Variation of Information *VI*

$$VI(C,C') = H(C) + H(C') - 2I(C,C')$$

Variation of Information to assess clustering distance (Meilă, 2007).

- VI can be normalized [0,1]
  - for same data sets:

$$V(C,C') = \frac{1}{\log n} VI(C,C')$$

  - for same number of clusters.

$$V_{k^*}(C,C') = \frac{1}{2\log K^*} VI(C,C')$$

Hypothesis

H1: Community Detection in informal communication networks is possible without semantic analysis and the graph that represents this informal communication network holds enough information to characterize it’s communities and it’s hierarchies.

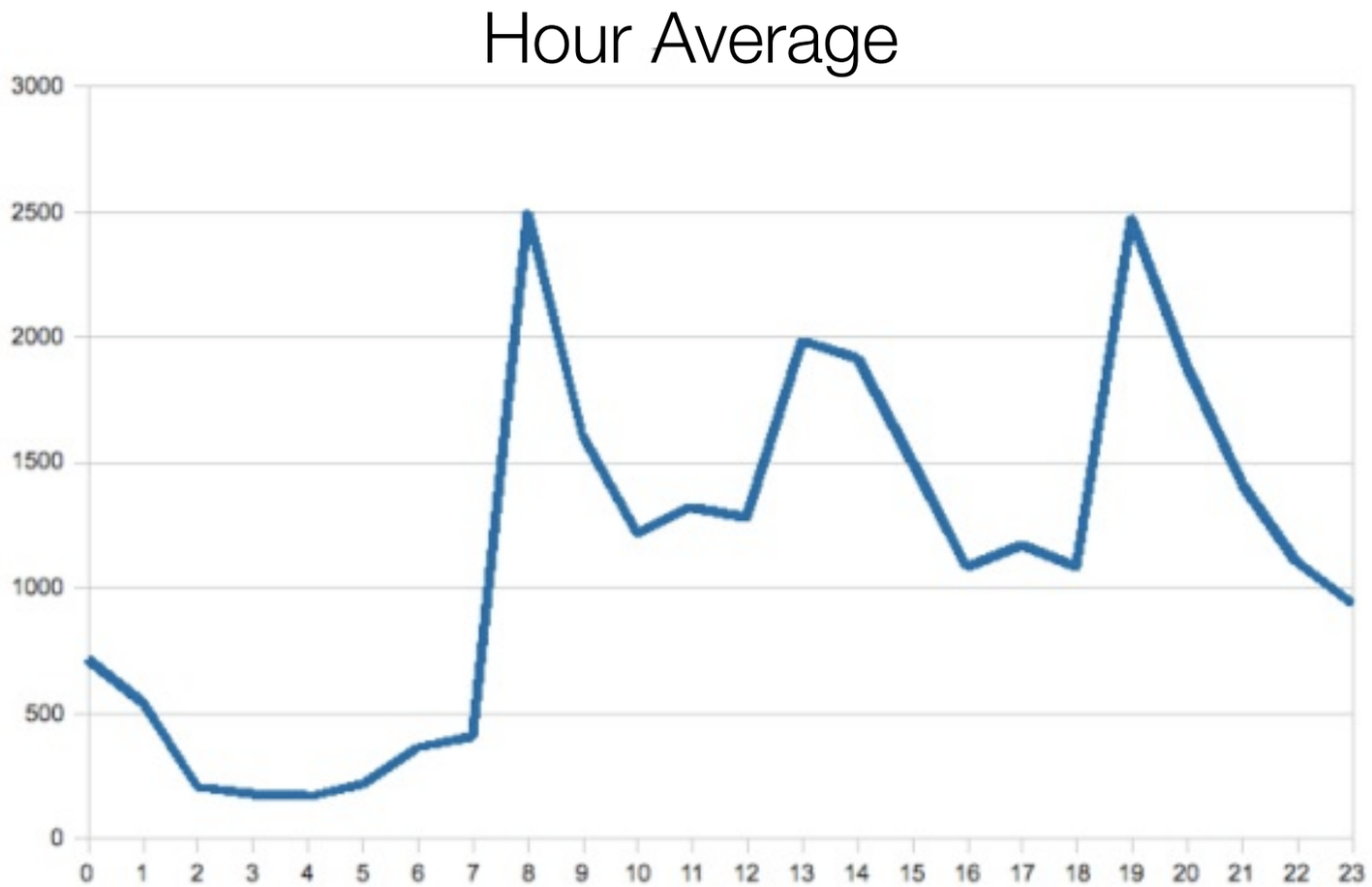
Case Study - Characterization

ISCTE Email Service

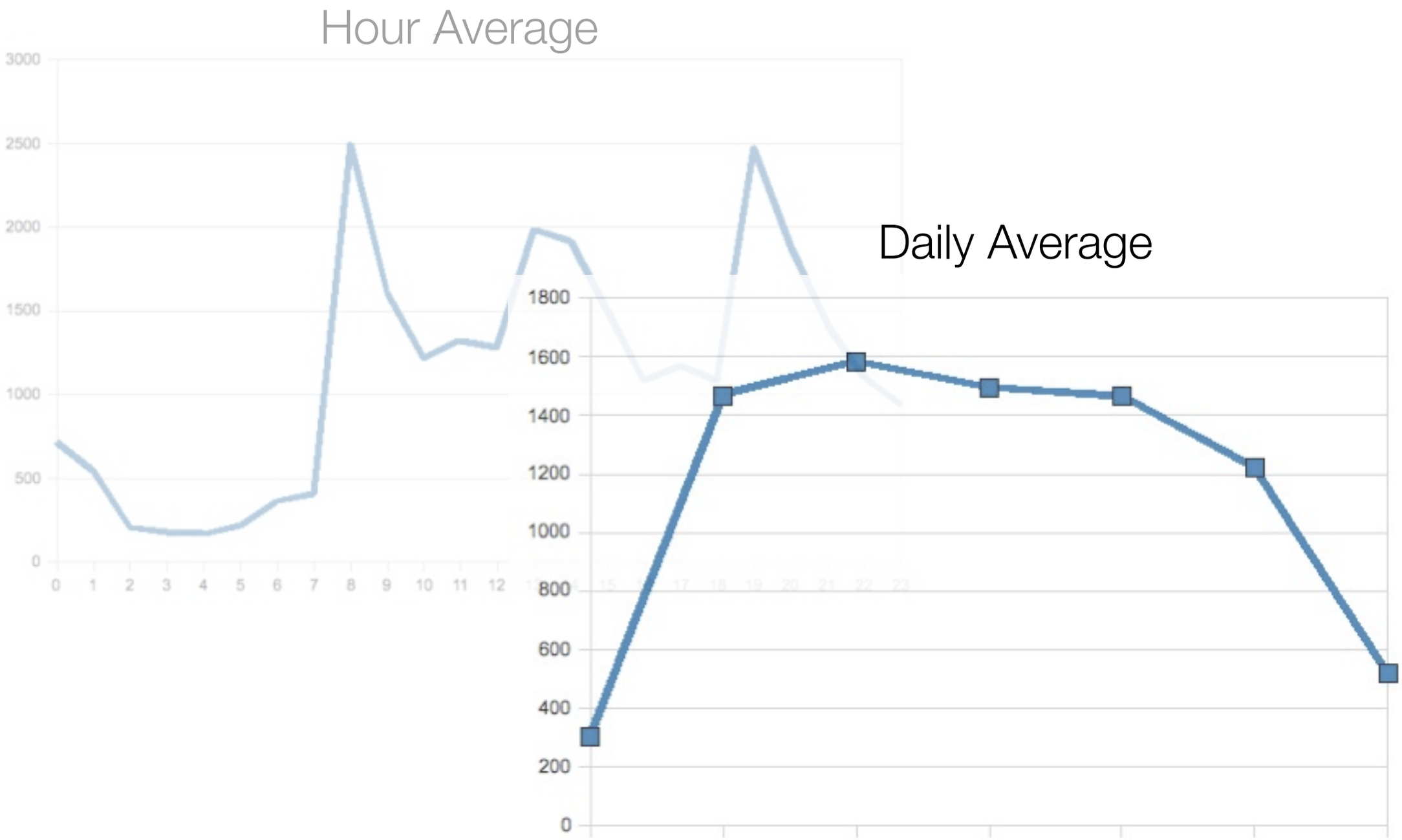
- 62 days
- 1 670 313 total messages / 242 544 processed
- 4 235 349 distinct recipients
- After Filtering and Cleaning data:
  - 3 Networks: Teachers, Students, Employees

	# members	# total	perc.
Teachers	395	1153	34%
Students	279	11698	2,4%
Employees	197	426	46%

ISCTE Email Service



ISCTE Email Service

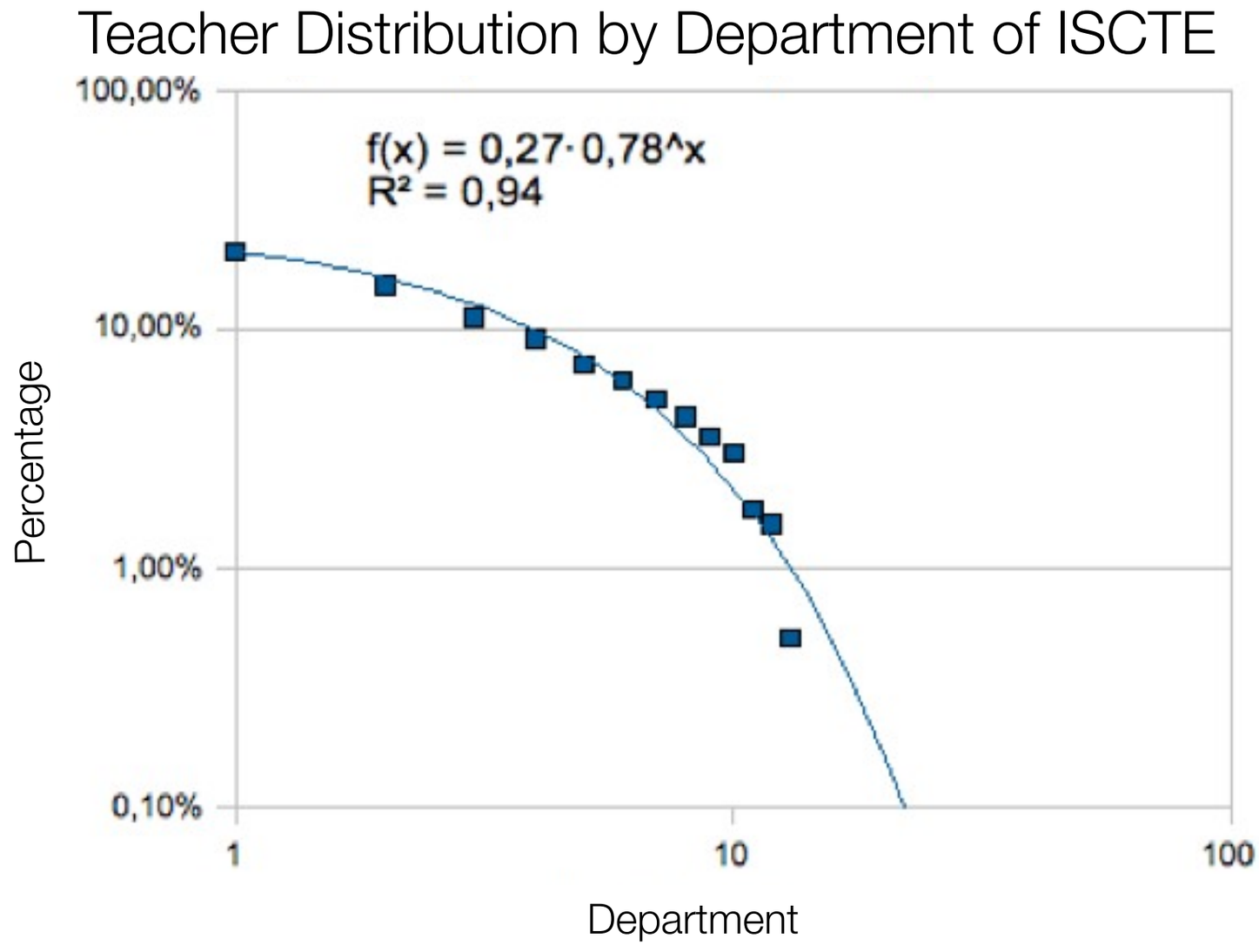


Characterization

Teacher Distribution by Department of ISCTE

Department	n.ºTeachers	Percentage
DCTI	83	21,01%
DMQ	60	15,19%
DE	44	11,14%
DS	36	9,11%
DCG	28	7,09%
DA	24	6,08%
DPSO	20	5,06%
DH	17	4,30%
SAAU	14	3,54%
DF	12	3,04%
SAD	7	1,77%
DC	6	1,52%
ACEA	2	0,51%
não ident.	42	10,63%
Total	395	100,00%

Characterization



Girvan-Newman Communities

- Identified 14 Communities
- Several small groups that in reality might be aggregated into bigger communities
- Modularity Q= 0,588

Identified Communities via the Girvan-Newman algorithm

Grupo	1	Grupo	2	Grupo	3
DS	3	DMQ	54	DPSO	19
Outros	1	DCG	28	Outros	9
		Outros	12	DS	8
		DF	11	DE	1
		DC	5		
		DS	2		
		DPSO	1		
Total	4		113		37
Grupo	4	Grupo	5	Grupo	6
DCTI	83	DS	17	DE	43
DMQ	5	Outros	10	Outros	4
DS	1	SAD	5	SAD	2
Outros	2	DMQ	1	DC	1
Total	91		33		50
Grupo	7	Grupo	8	Grupo	9
DS	2	DA	24	SAAU	6
Outros	1	DH	17		
		SAAU	2		
		Outros	1		
Total	3		44		6
Grupo	10	Grupo	11	Grupo	12
DS	2	DF	1	ACEA	2
Outros	2				
Total	4		1		2
Grupo	13	Grupo	14		
SAAU	6	DS	1		
Total	6		1		

Clauset-Newman-Moore Communities

- Identified 7 communities
- Groups are more heterogeneous
- Modularity Q = 0,585
- Both GN and CNM had Q>0,3

Identified Communities via the Clauset-Newman-Moore algorithm

Grupo	1	Grupo	2	Grupo	3
DCTI	80	DPSO	19	DCG	25
DMQ	5	DCTI	3	SAD	5
DCG	2	DS	5	DF	7
Outros	2	DE	1	DMQ	3
		Outros	9	DC	5
				Outros	10
Total	89	Total	37	Total	55

Grupo	4	Grupo	5	Grupo	6
DMQ	52	DE	42	DH	17
DS	30	DC	1	DA	24
SAD	2	Outros	4	SAAU	2
DE	1			DF	5
DPSO	1			DS	1
ACEA	2			DCG	1
Outros	14			Outros	3
Total	102	Total	47	Total	53

Grupo	7
SAAU	12
Total	12

Variation of Information

$H(\text{ISCTE})$	2.346
$H(\text{Girvan-Newman})$	1.945
I	1.474
VI	1.342
V	0.224

$V^{random} = 0.833$

$H(\text{ISCTE})$	2.346
$H(\text{Clauset-Newman-Moore})$	1.811
I	1.372
VI	1.413
V	0.236

$V^{random} = 0.718$

$H(\text{Girvan-Neman})$	1.945
$H(\text{Clauset-Newman-Moore})$	1.811
I	1.390
VI	0.976
V	0.163

k-core analysis

- The *k*-core analysis isn't homogeneous.

- DCTI, DMQ or DE have high percentages of teachers very well connected while other like DS are spread. Others like SAAU only have teachers in the peripheral *k*-cores.

Member distribution by *k-core*

	1	2	3	4	5	6	7	8	9
DCTI	9	3	7	9	5	2	1	6	43
DMQ	4	3		3	3	8	9	34	
DE	3	4	1	1		1	1	5	32
DS	9	4	8	3	3	8	7	5	1
DCG	2	1	3			1	7	19	
DPSO		1	1	1	2	3	9	7	1
DA			4	3		1	2	15	
DH	2	1	2	1	2	3	6		
SAAU	5	8		1					
DF	1		6	2				5	
SAD	1	2	1	2		2			
DC	3		2		2				
ACEA	1	1							1
Outros		1		3					

k-core analysis

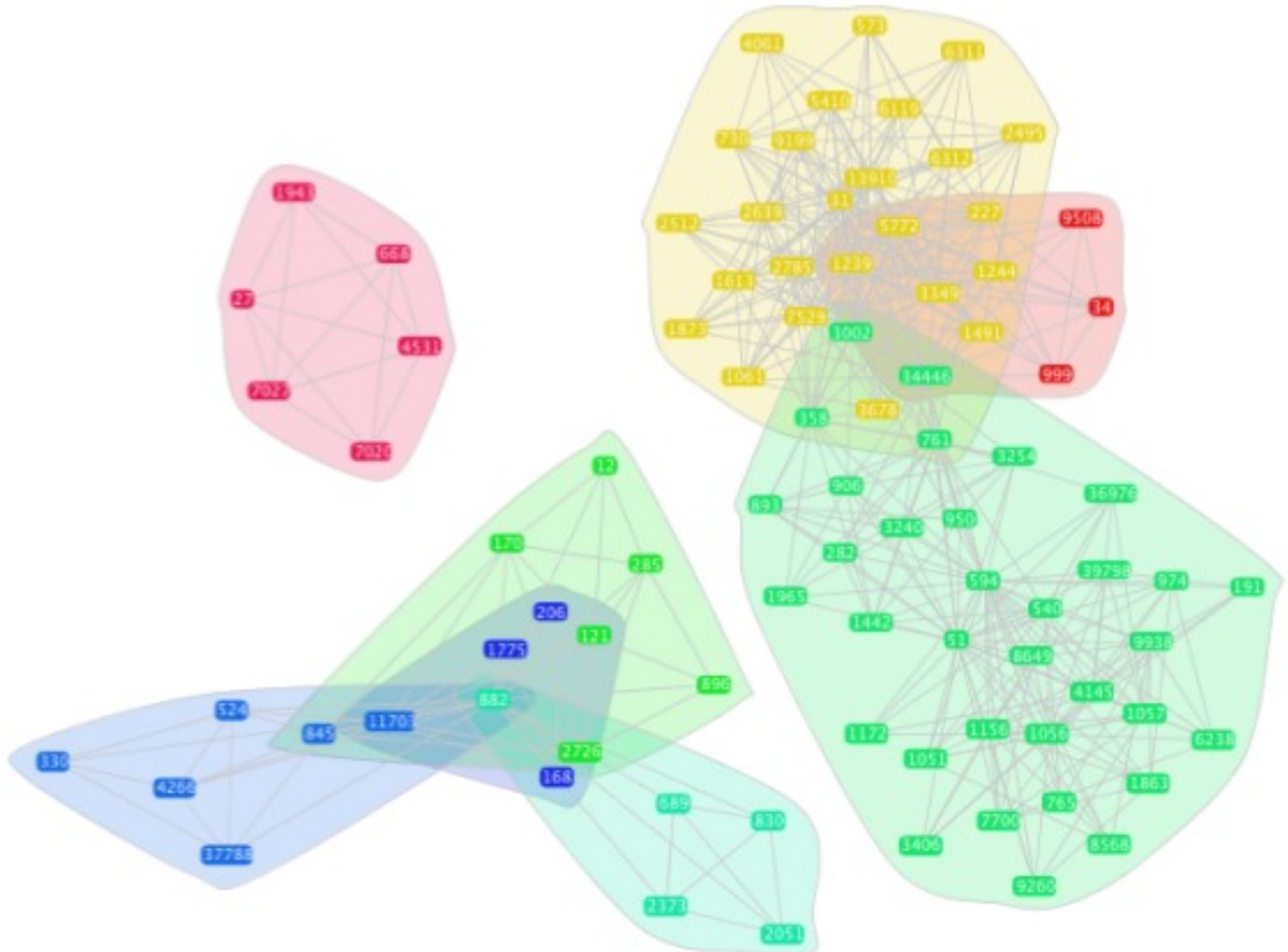
- The *k*-core analysis isn't homogeneous.
- DCTI, DMQ or DE have high percentages of teachers very well connected while other like DS are spread. Others like SAAU only have teachers in the peripheral *k*-cores.
- The number of members to each *k*-core (almost linear)

***k-core* Elements**

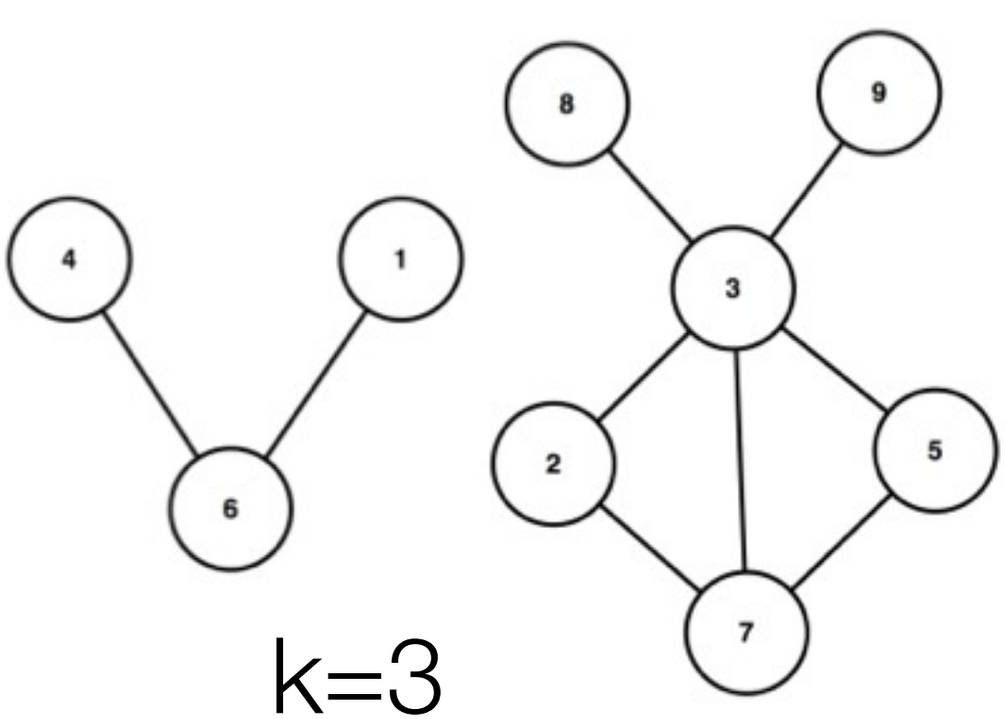
k-core	n.ºElements	*k-core* Total
1	40	395
2	29	355
3	35	326
4	29	291
5	17	262
6	29	245
7	42	216
8	96	174
9	78	78

Clique Percolation

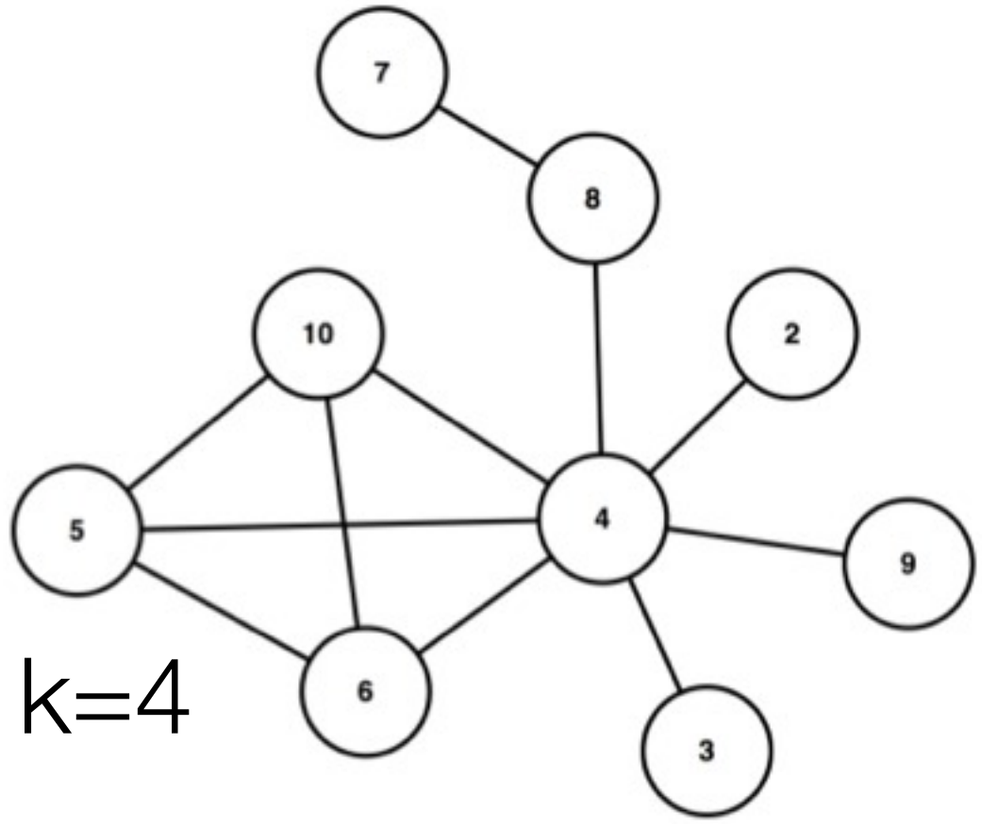
- Used clique percolation for k=<3... 7> (right pic k=5)
- 3 separate communities
- Overlapping is visible in 2 of the 3



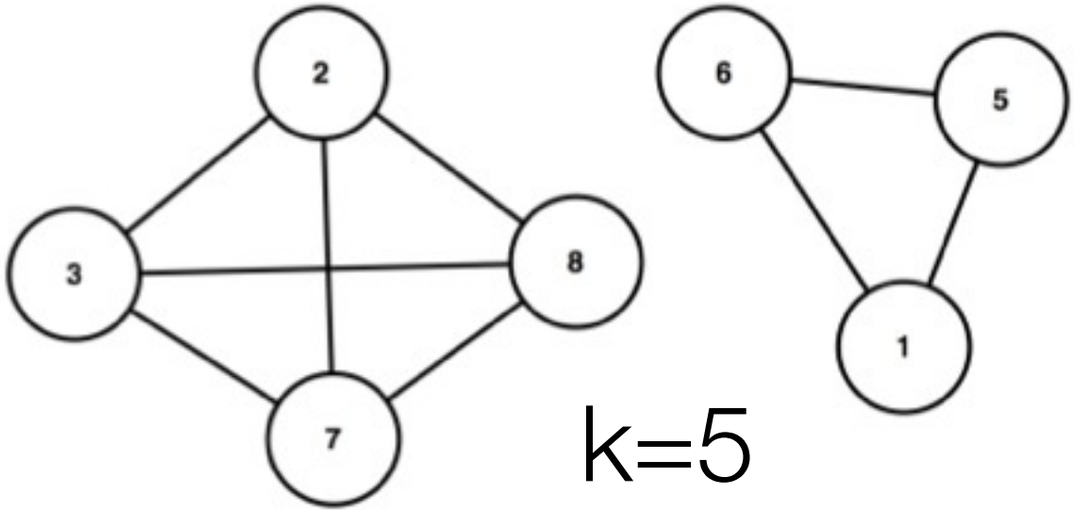
Clique Percolation



k=3



k=4



k=5

Analysis Comparison

Results Summary

Method	Ambit	Components	Groups	N.º of Classified
Girvan-Newman	Global	14	14	395
Clauset-Newman-Moore	Global	7	7	395
k-core	Local			
$k = 1$		1	-	395
$k = 2$		1	-	355
$k = 3$		1	-	326
$k = 4$		1	-	291
$k = 5$		1	-	262
$k = 6$		1	-	245
$k = 7$		1	-	216
$k = 8$		1	-	174
$k = 9$		1	-	78
Percolação de cliques	Local			
$k = 3$		2	9	300
$k = 4$		2	10	195
$k = 5$		3	8	105
$k = 6$		2	2	35
$k = 7$		2	2	19

Conclusions

Conclusions and Contributions

- The latent structure of the network has enough information to characterize the communities and hierarchies present in the network

- Global algorithms based on global properties are able to identify communities without looking into the semantic content.

- Through k-core analysis hierarchies are revealed, showing asymmetric group compositions

- The informal communication network transverses the realm of departments and this was shown via clique percolation theory.

Conclusions and Contributions

- Future work will include the application of information theory measures to directed and weighted networks and also to the field of “tagged networks”

- Future work should also investigate the several layers / networks where the self is included as the individual is affected by the interplay of several informal communications networks.

- In teaching environments this might be useful to understand how communication processes affect and help learning.

References

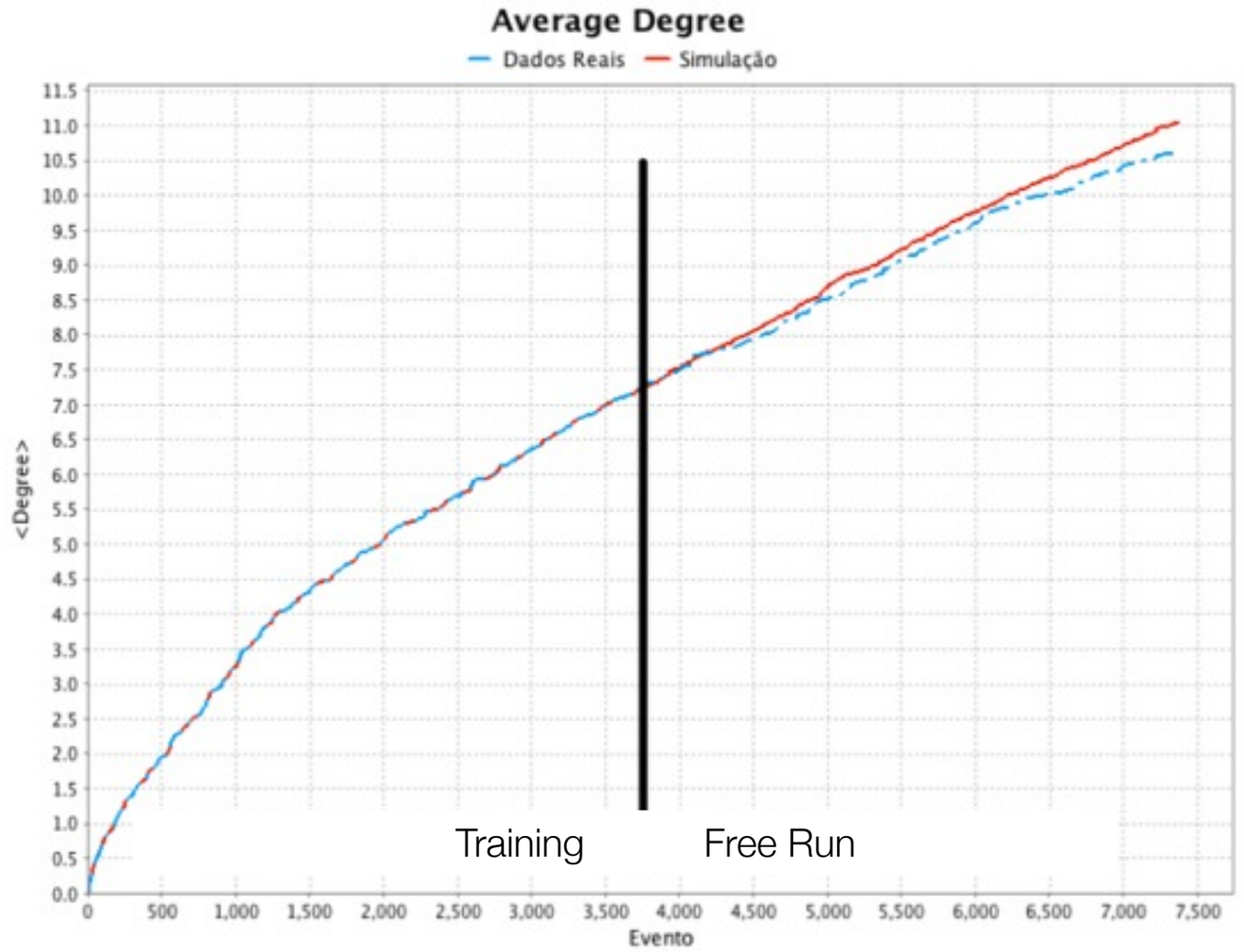
- Euler, L. (1741). Solvtio Problematis Ad Geometriam Sitvs Pertinentis. Commentarii academiae scientiarum Petropolitanae, 8(53), 128-140.
- Girvan, M., & Newman, M. E. J. (2001, December). Community structure in social and biological networks. cond-mat/0112110, from http://arxiv.org/abs/cond-mat/0112110
- Clauset, A., Newman, M. E. J., & Moore, C. (2004, August). Finding community structure in very large networks. cond-mat/0408187, from http://arxiv.org/abs/cond-mat/0408187
- Dorogovtsev, S. N., Goltsev, A. V., & Mendes, J. F. F. (2005, setembro). k-core organization of complex networks. cond-mat/0509102, from http://arxiv.org/abs/cond-mat/0509102
- Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435, 814-818.
- Meilă, M. (2007). Comparing clusterings---an information based distance. J. Multivar. Anal., 98(5), 873--895.

Thank You

david.rodrigues@gmail.com

CIUCEU MABS Model.

- CIUCEU = **C**omunicação **I**nformal entre **U**tilizadores de **C**orreio **E**lectrónico **U**niversitário.
- Use Real Data
- “Social Neighborhood”
  - High Transitivity
- Assortative Mixing
  - Preferential Attachment



CIUCEU MABS Model.

- Teachers have an array of Probabilities of contacts updated at the end of each step according to the number of historic contacts
- Assortative Mixing
- Transitivity

$$pc_{i,j} = \frac{\frac{p_j}{1+|k_i-k_j|}}{\sum_j \frac{p_j}{1+|k_i-k_j|}}$$

$$p^*_{i,l} = \frac{1}{d^a} p_{j,l}$$

$$p_{i,l} = \frac{\sum_{cam.geod} p^*_{i,l}}{\sum_k \sum_{cam.geod} p^*_{i,l}}$$

CIUCEU MABS Model.

- Teachers have an array of Probabilities of contacts updated at the end of each step according to the number of historic contacts

- Assortative Mixing

- Transitivity