Novel Measures on Directed Graphs and Applications to Large-Scale Within-Network Classification

Amin Mantrach
Phd public defense
IRIDIA
Université Libre de Bruxelles
November 23, 2010

Contributions:

(1) ▸ First Contribution : Novels Measures on Directed Graphs

A novel **centrality** measure: **betweenness**
A novel **relatedness** measure: **covariance**

Amin Mantrach, Luh Yen, Jerome Callut, Kevin Francoisse, Masashi Shimbo, and Marco Saerens. The sum-over-paths covariance kernel: A novel covariance measure between nodes of a directed graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:11121126, June, 2010.

(2) ▸ Second Contribution : Applications to Large-Scale Within-Network Classification

Nodes classification on **Large-Scale**, Sparse, Directed Graphs.
A novel data set collected: The U.S. Patents Citation Network

Amin Mantrach, Nicolas van Zeebroeck, Pascal Francq, Masashi Shimbo, Hugues Bersini and Marco Saerens. Semi-supervised Classification and Betweenness Computation on Large, Sparse, Directed Graphs, *to appear in Pattern Recognition*, PR-D-09-01097R.

(3) ▸ Third Contribution : Applications to Large-Scale Within-Network Classification

Combining **citation-based** graphs with **content-based** data

▸ Conclusions & Perspectives

Network Data - Some Popular Web Sites

www.google.com

More than **30 billion** of pages

www.facebook.com

More than **500 millions** of users - Average user has 30 friends

en.wikipedia.com

More than **3.5 million** of articles

Network Data

- Web pages are pointing to other pages

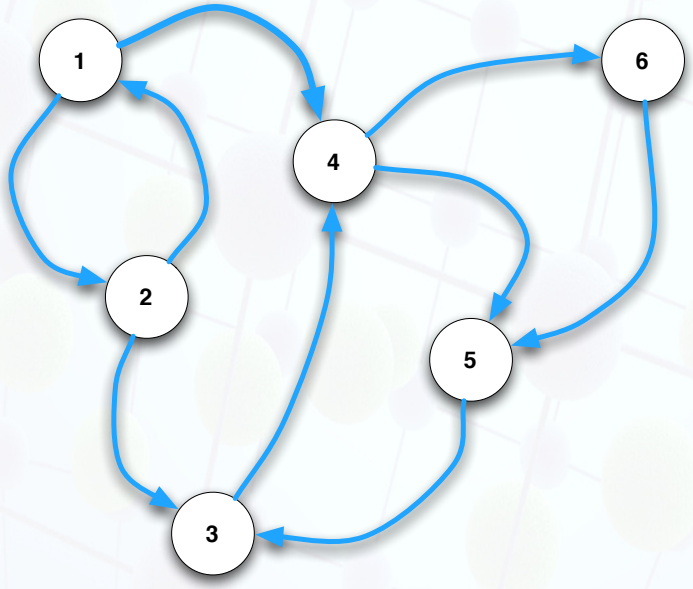


Figure: **Web pages forming a directed graph.**

Network Data

- On facebook users are linked through friendship relation

Figure: **Users forming an undirected Graph.**

Why analyzing networks is important: An Example

Recommander à des amis

This page is run by Organizing for America, the grassroots organization for President Obama's agenda for change. OFA is a project of the Democratic National Committee. To visit the White House Facebook page, go to facebook.com/WhiteHouse.

Informations

Poste actuel
Bureau :
President of the United States

Barack Obama J'aime

Mur | Infos | Photos | OFA Store | Vidéo

Barack Obama The DREAM Act would provide a path to citizenship for undocumented youth who are willing to work for a college degree or serve in our armed forces. Add your name to show that you support this important legislation.

Stand with the President to Pass the DREAM Act
my.barackobama.com
President Obama believes that the DREAM Act should be law. Adding your name will be a strong signal to lawmakers that the American public supports this important step forward.

Hier, à 00:51 · Afficher le feedback (25 754) · Partager

Barack Obama Because of the tough decisions we made, the American auto industry—an industry that's been the proud symbol of America's manufacturing might for a century and that helped to build our middle class—is once again on the rise.

President Obama on GM: "One of the Toughest Tales" Becoming a "Success Story"
www.youtube.com
President Obama made a statement from the White House about General Motors' relaunch as a public company.

vendredi, à 23:59 · Afficher le feedback (11 833) · Partager

Why analyzing networks is important: An Example

Challenges

- What are the more **central** actors, i.e.: persons, web pages, wiki articles, etc.

Study and analyze the network

$\rightarrow$ Let us introduce the **IRIDIA** social network

The Gang of Thirty-Five

Bersini, Hugues – Dorigo, Marco – Birattari, Mauro – Sttzle, Thomas – Saerens, Marco – Decreton, Muriel – Coletta, Alain – De Beule, Joachim – Lpez-Ibez, Manuel – Marchal, Bruno – O'Grady, Rehan – Scheidler, Alexander – Trianni, Vito – Turgut, Ali Emre – Van Zeebroeck, Nicolas – Venet, David – Walker, Nick – Weiss Solis, David – Abbaci-Gaultier, Faza – bin Hussin, Mohamed Saifullah – Brambilla, Manuele – Brutschy, Arne – Campo, Alexandre – Decugnire, Antal Dubois-Lacoste, Jrmie – Ferrante, Eliseo – Lenne, Renaud – Liao, Tianjun – Mantrach, Amin – Mathews, Nithin – Montes de Oca, Marco – Oliveira, Sabrina – Pinciroli, Carlo – Pini, Giovanni – Stranieri, Alessandro – Yuan, Zhi Eric – Duqu, Robin – Benedettini, Stefano – Piscopo, Carlotta – Roli, Andrea

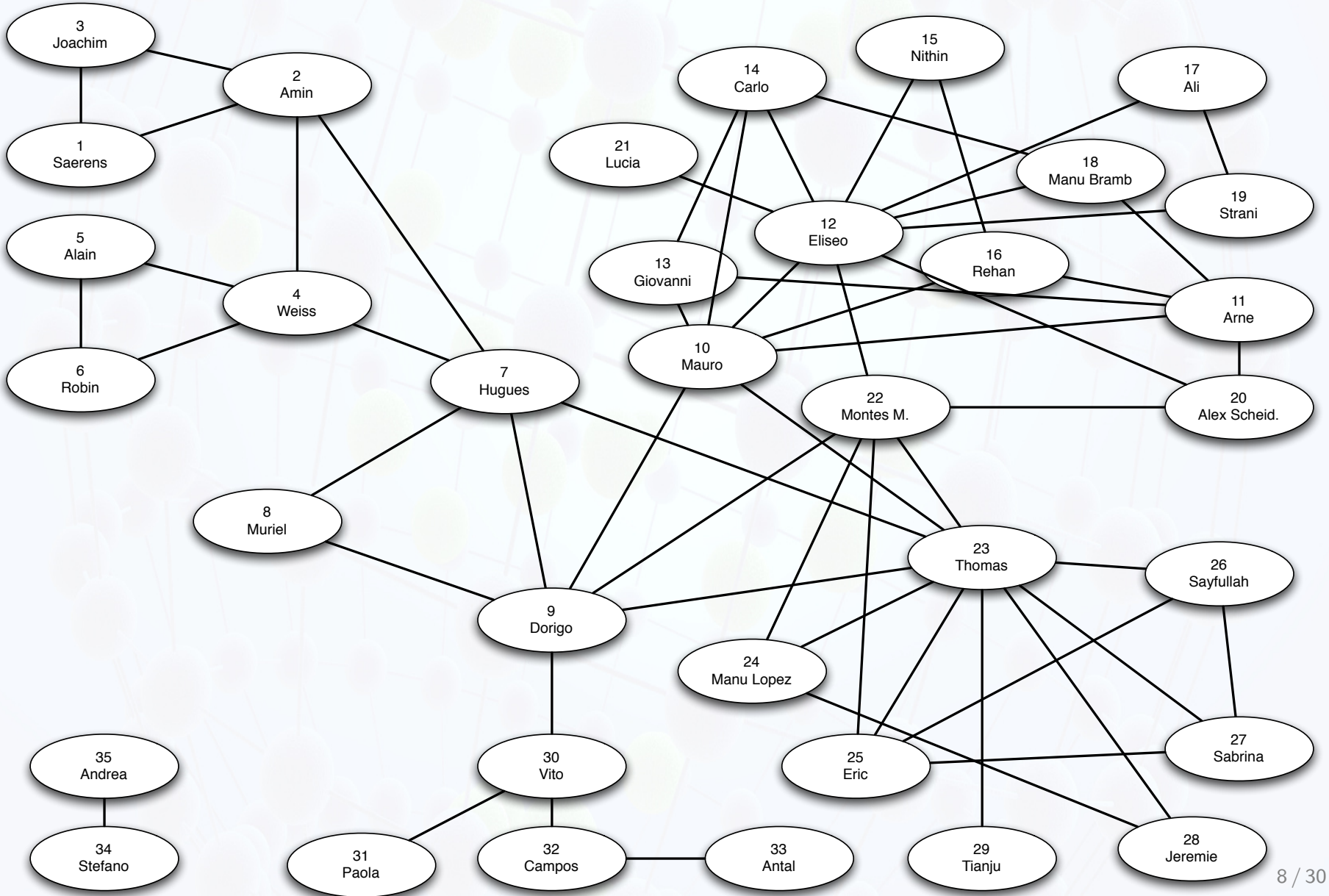
The Gang of Thirty-Five

- To build the graph we collected for each researcher a **list of the persons** with which he has the **strongest** interactions
- Finally, we keep a (undirected) link between two persons in case of **mutual citation**

Link Inference Example

- Hugue's list : Dorigo, Weiss, **Amin**, Thomas, Muriel
- Amin's list : **Hugues**, Saerens, Joachim
- Hugues $\longleftrightarrow$ Amin

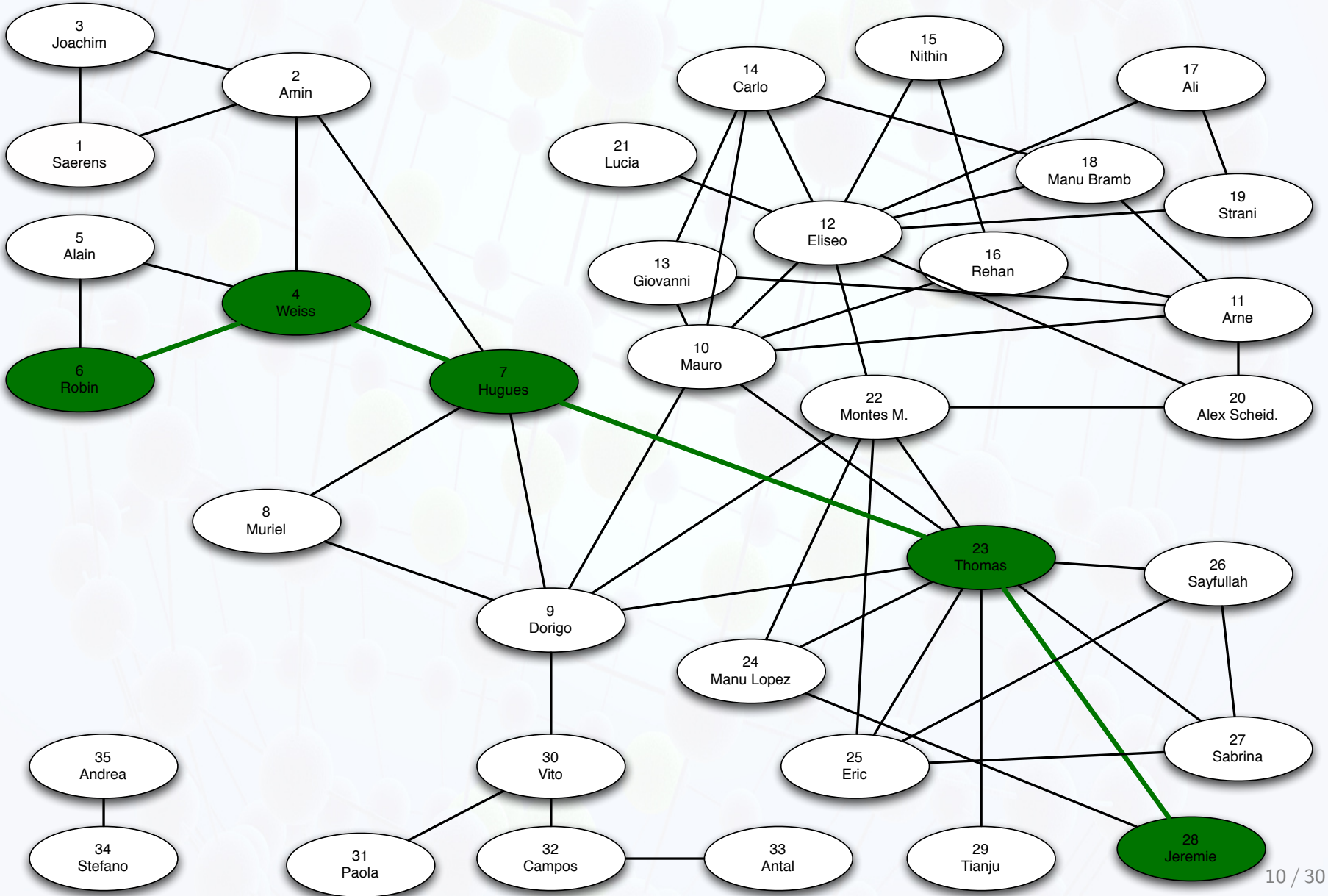
The Gang of Thirty-Five



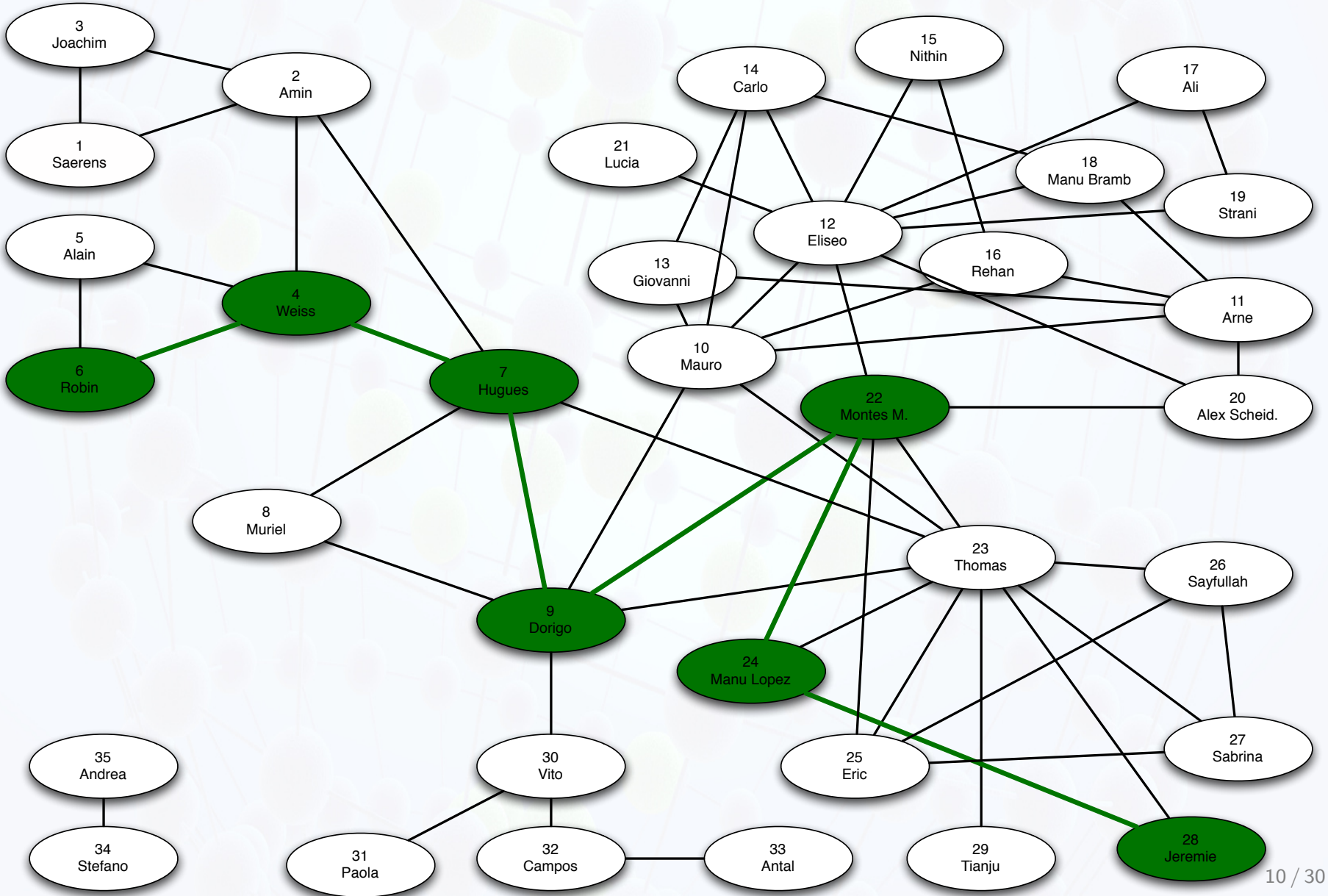
Analyze the centrality

- One measure to analyze the centrality is through the **betweenness**
- The all paths betweenness, of Newman, consists in considering **all possible paths** in the graph.
- And then compute the **average number of times** a node appears on the paths.

Different possible paths



Different possible paths



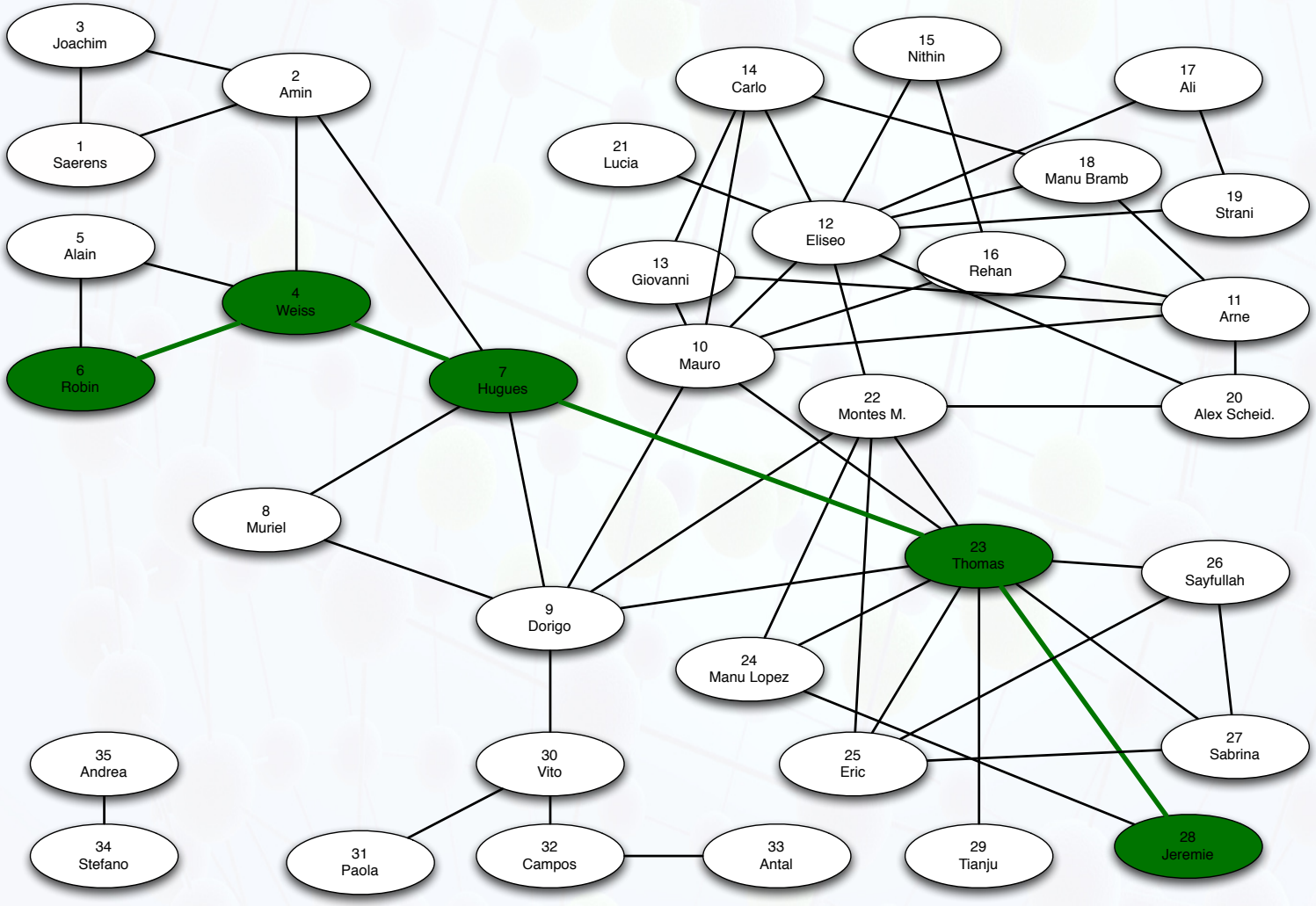
Betweenness

Therefore, we can rank the nodes according to the **all paths betweenness**:

All paths

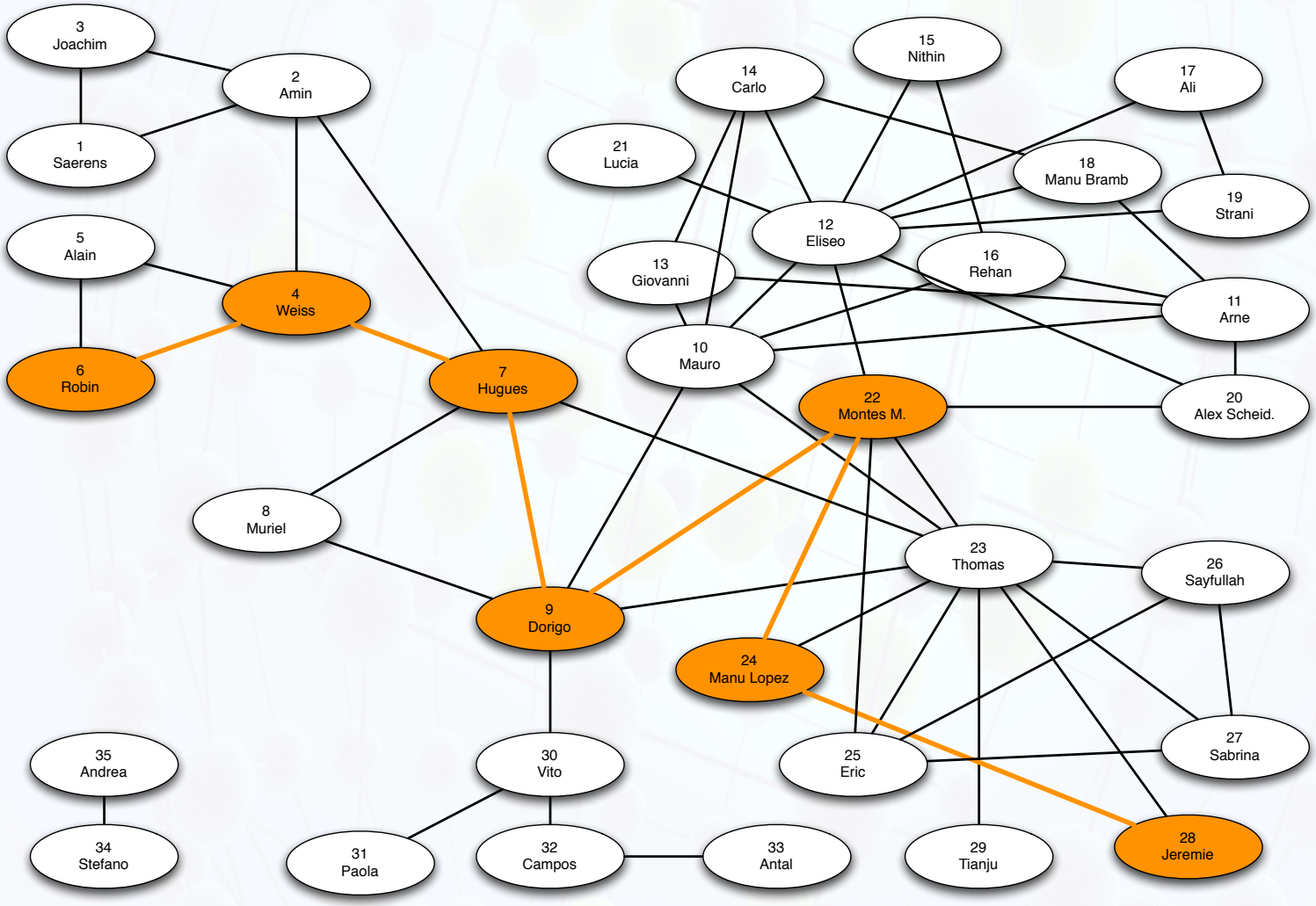
1. Thomas: 8,4%
2. Eliseo: 7.6%
3. Mauro: 5.9%
4. Marco Dorigo & Montes: 5%
5. Hugues & Arnee: 4.2%
6. Amin & Weiss: 3.4%

- However, we may prefer to **decrease the importance** of too long paths, by biasing the measure in **favor of short paths**.



Favor short paths

- However, we may prefer to **decrease the importance** of too long paths, by biasing the measure in **favor of short paths**.



But consider also (with less weight) longer paths

Other Rankings

Tradeoff

1. Eliseo: 8.8%
2. Thomas: 8.5%
3. Mauro: 4.8%
4. Marco Dorigo: 4.5%
5. Marco Montes: 4.3%
6. Hugues: 4.1%

All paths

1. Thomas: 8,4%
2. Eliseo: 7.6%
3. Mauro: 5.9%
4. Marco Dorigo & Montes: 5%
5. Hugues & Arnee: 4.2%
6. Amin & Weiss: 3.4%

Other Rankings

Tradeoff

1. Eliseo: 8.8%
2. Thomas: 8.5%
3. Mauro: 4.8%
4. Marco Dorigo: 4.5%
5. Marco Montes: 4.3%
6. Hugues: 4.1%

All paths

1. Thomas: 8,4%
2. Eliseo: 7.6%
3. Mauro: 5.9%
4. Marco Dorigo & Montes: 5%
5. Hugues & Arnee: 4.2%
6. Amin & Weiss: 3.4%

$\rightarrow$Tradeoff exploration / exploitation

We can also bias completely the measure by **considering only the shortest-paths**.

All the spectrum...

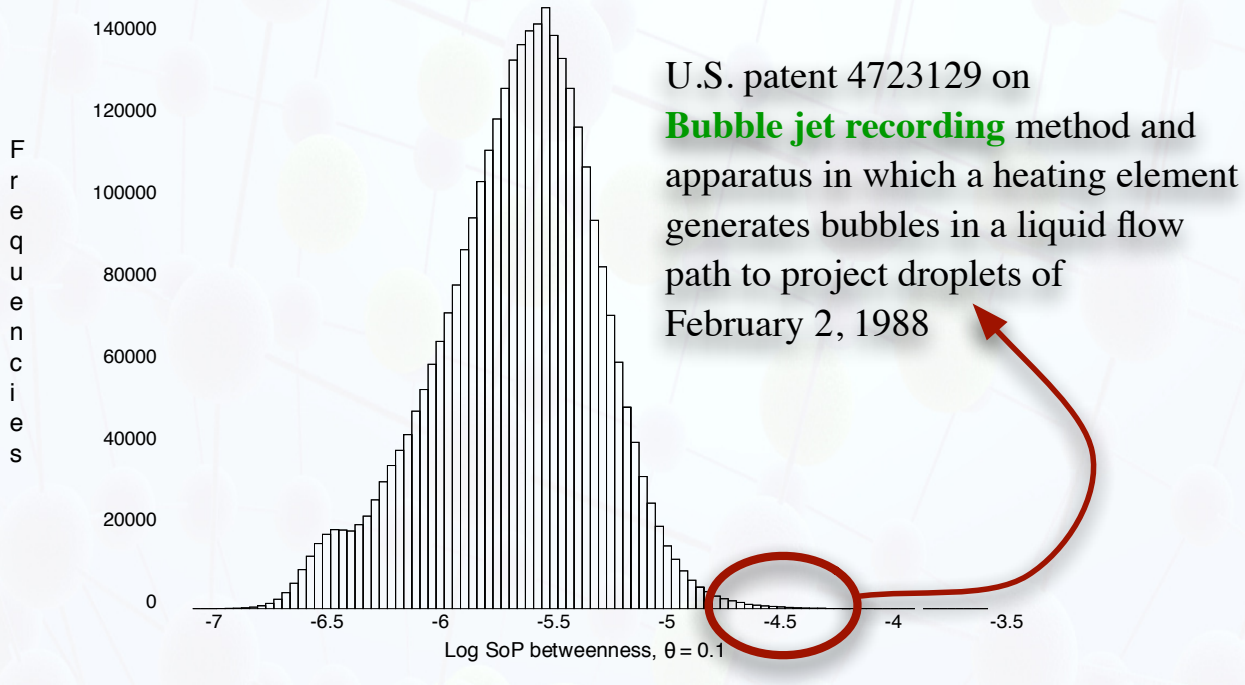
- A novel betweenness measure, with a temperature parameter giving the possibility to set the **tradeoff between exploration – exploitation**.

Shortest paths	Tradeoff	All paths
1. Eliseo: 10%	1. Eliseo: 8.8%	1. Thomas: 8,4%
2. Thomas: 9.8%	2. Thomas: 8.5%	2. Eliseo: 7.6%
3. Vito: 4.8%	3. Mauro: 4.8%	3. Mauro: 5.9%
4. Mauro: 4,3%	4. Marco Dorigo: 4.5%	4. Dorigo & Montes: 5%
5. Arne: 4.2%	5. Marco Montes: 4.3%	5. Hugues: 4.2%
6. Amin & Weiss: 4.14%	6. Hugues: 4.1%	6. Arnee: 4.2%

Computation of the Betweenness

A Novel Data Set for the Community: The U.S. Patents Citation Network

- around **3M** of patents granted between 1963 and 2002
- **38M** of (cited - citing) links - **6** broad areas (technological classes)



Community Detection

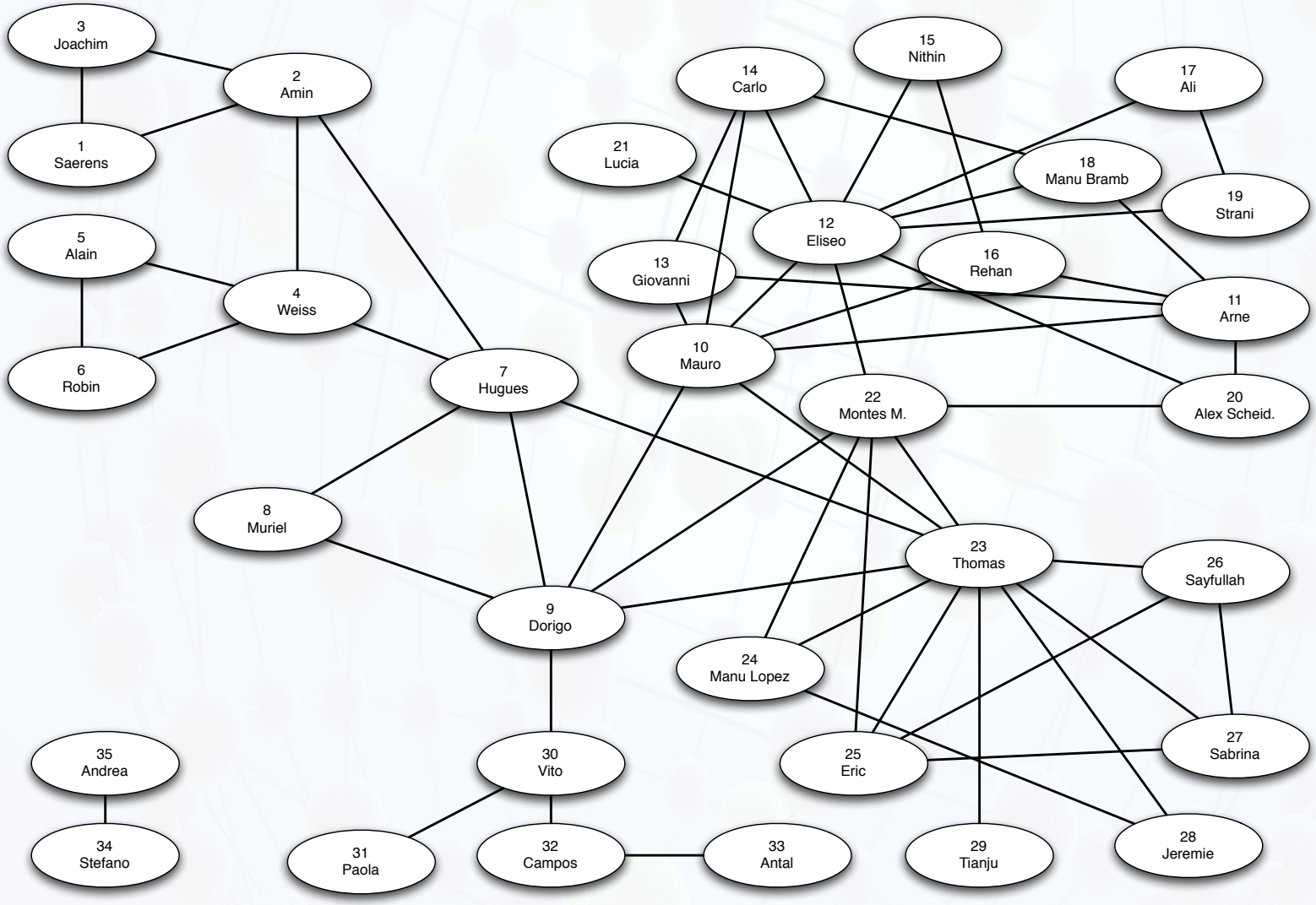
- Another important application in graph analysis consists in **detecting communities** (i.e., dense webs).
- Therefore we need to assess the **similarity** between different pair of nodes in the graph.
- In this thesis, we introduce, based one the same framework, a novel similarity measure between two nodes (i.e. entities) in a graph.

Novel Correlation Between Nodes in Graph

Two nodes are **highly correlated** if they often appear together in the same – preferably short – path.

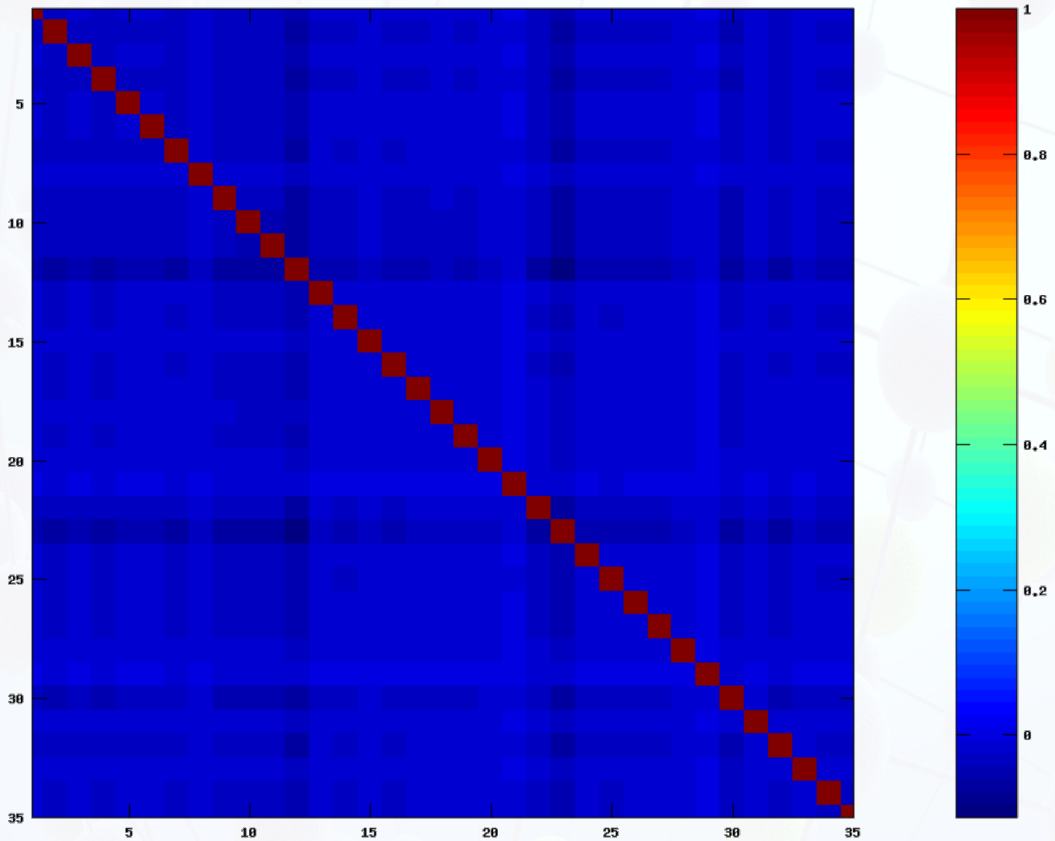
Community Detection

The IRIDIA members network:



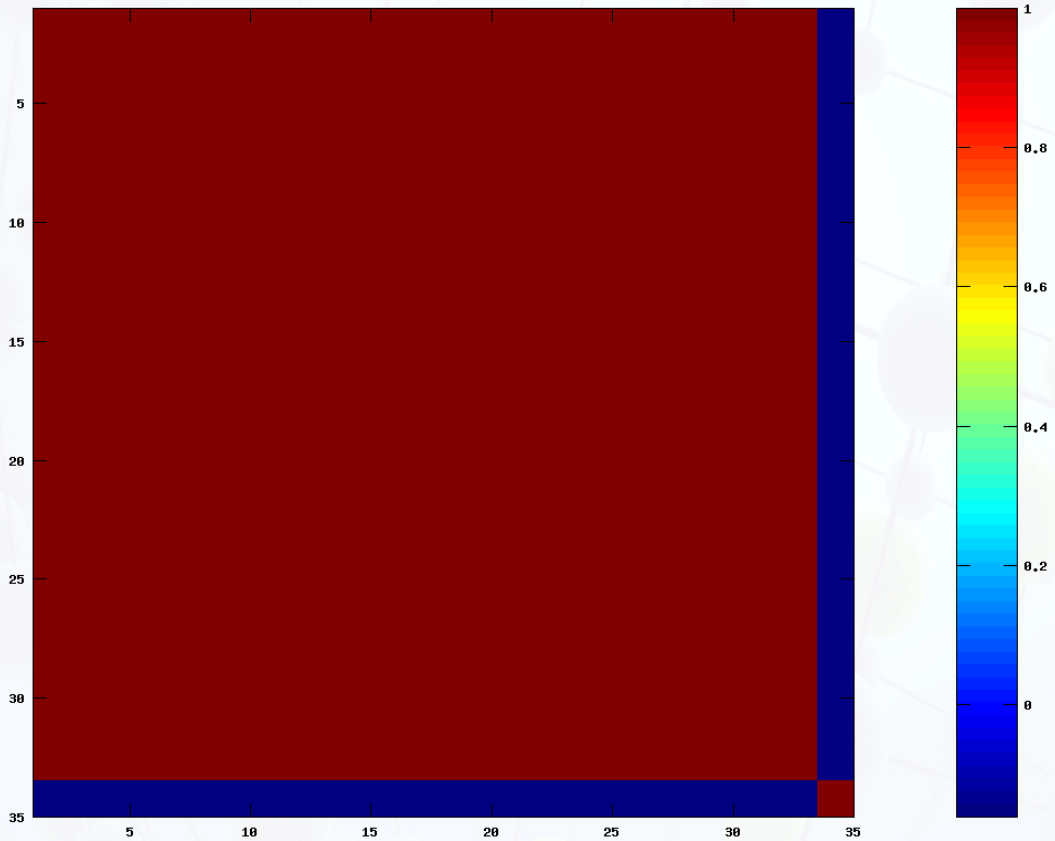
Community Detection

Taking only **short paths**, two nodes **rarely appear together** on the same path just a few time → To low proportion



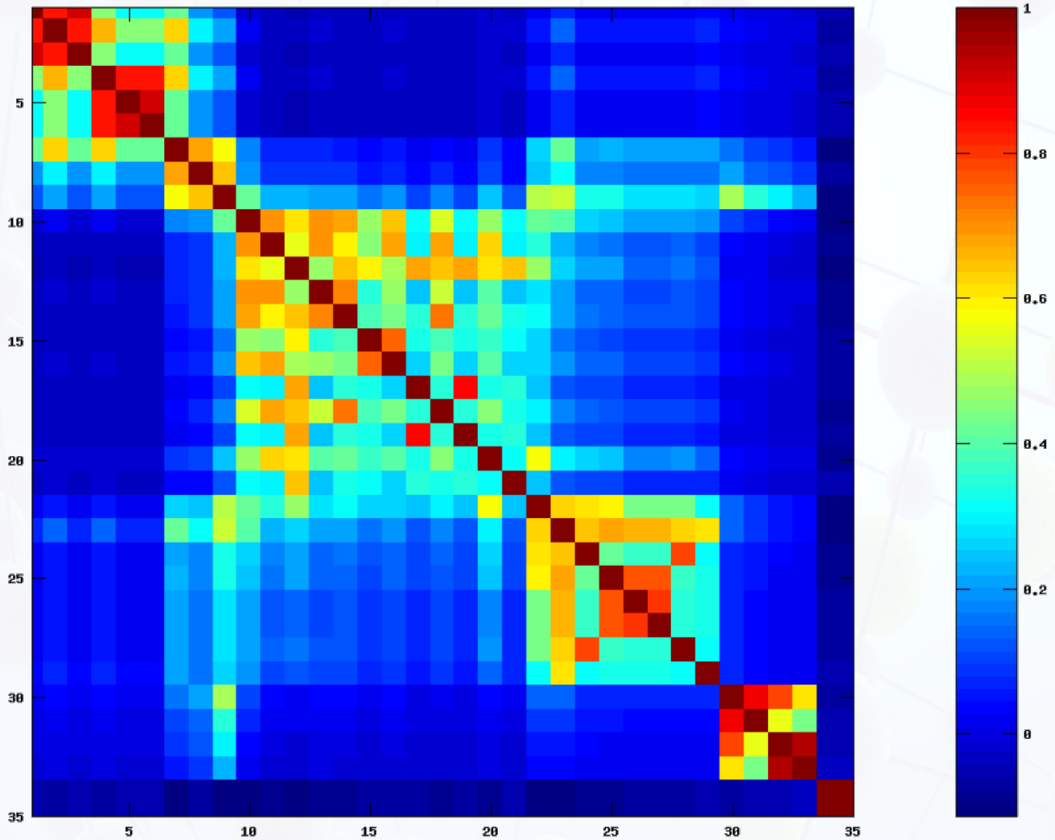
Community Detection

Taking **all paths**, two nodes appear together pratically always → Too high proportion

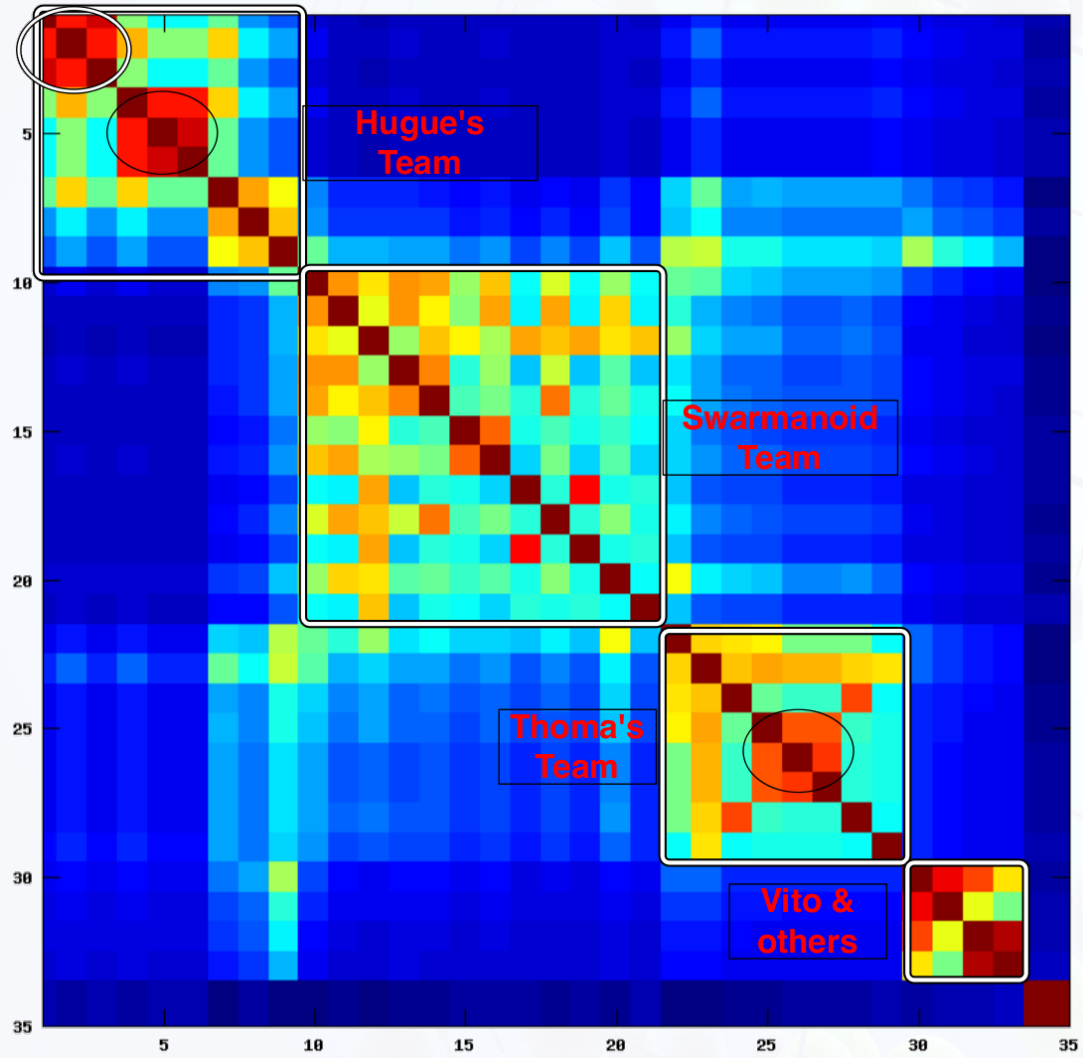


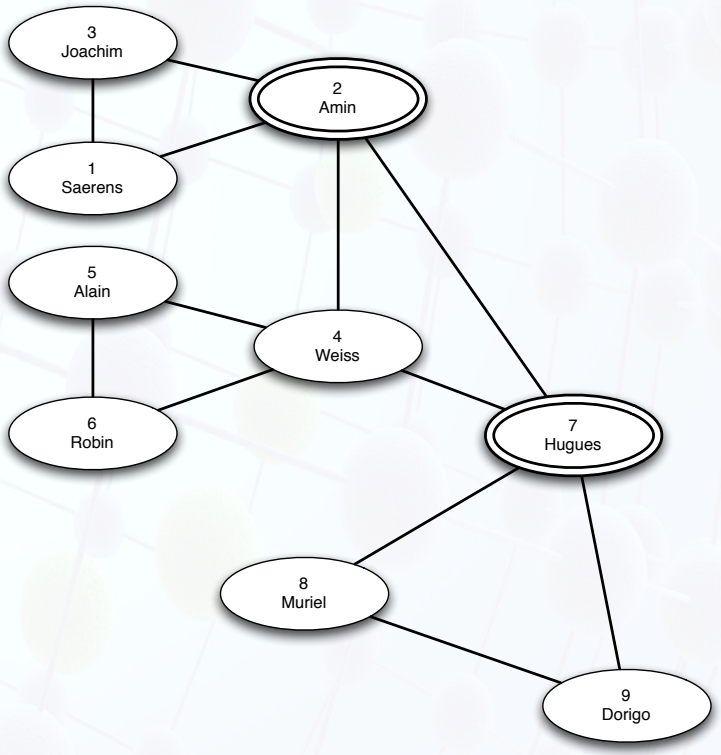
Community Detection

By choosing a good **tradeoff between exploration and exploitation**, we can obtain the following

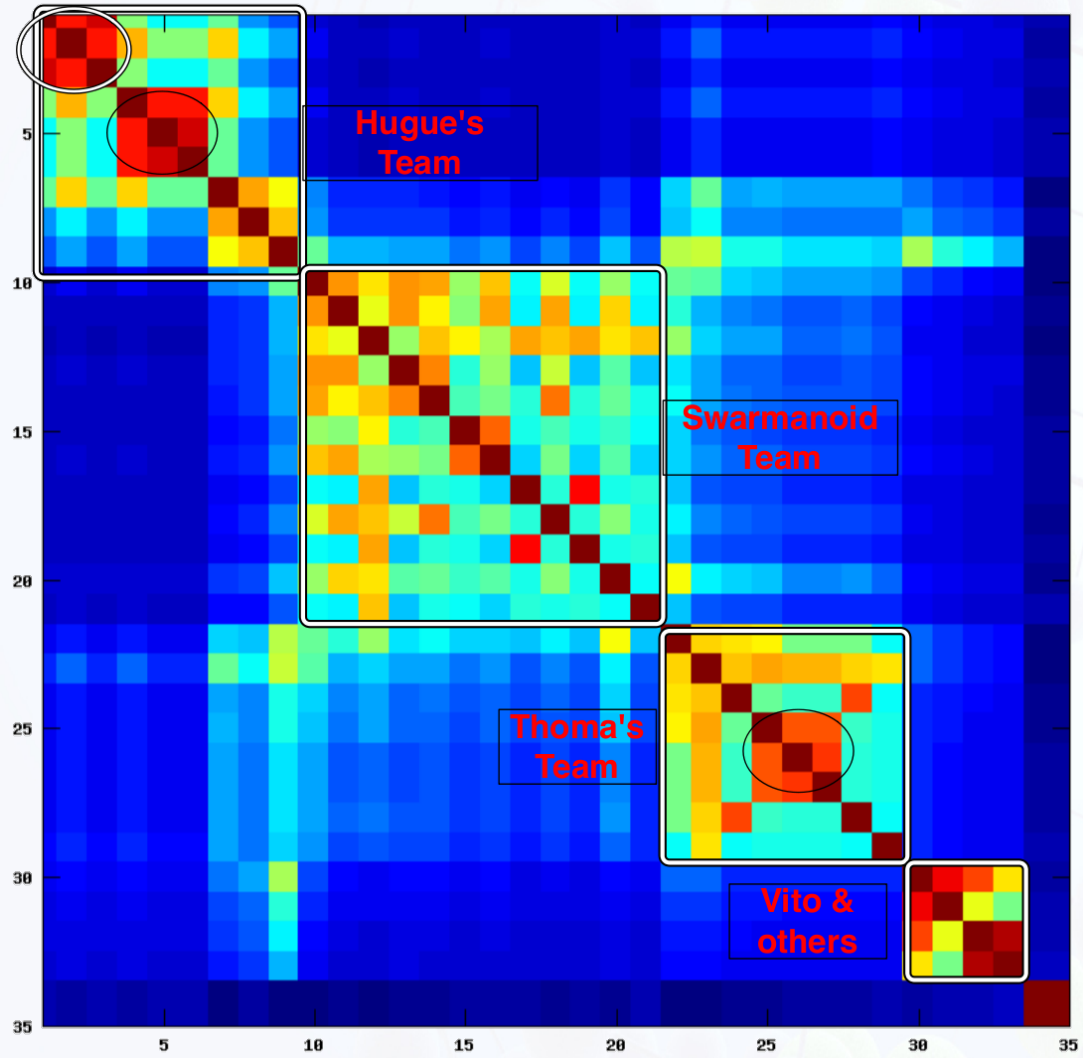


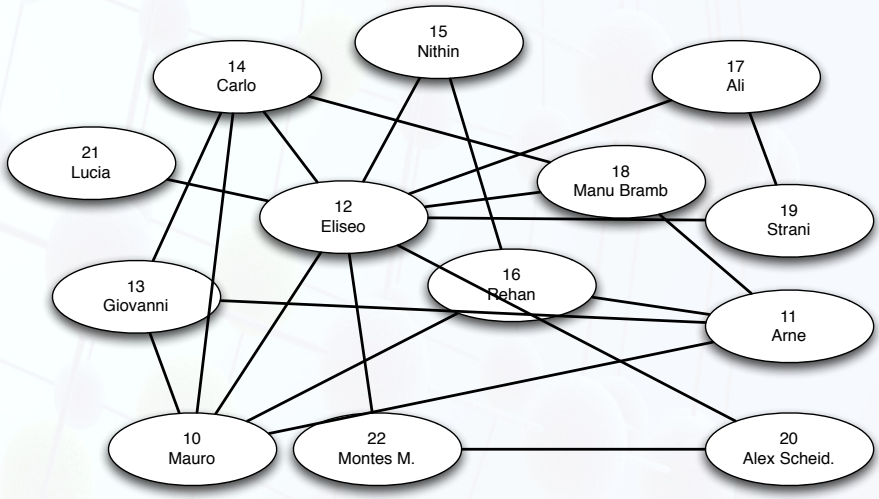
Community Detection



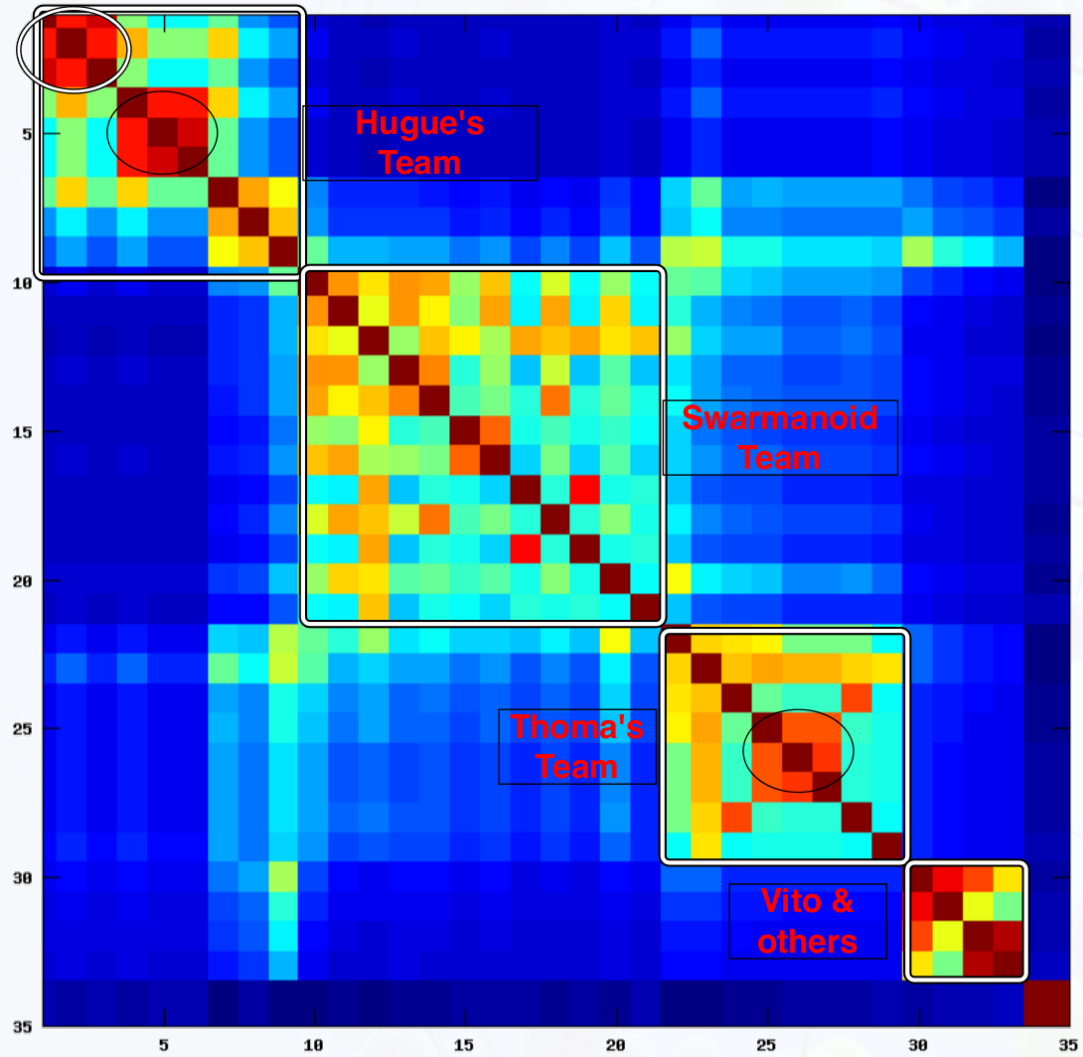


Community Detection

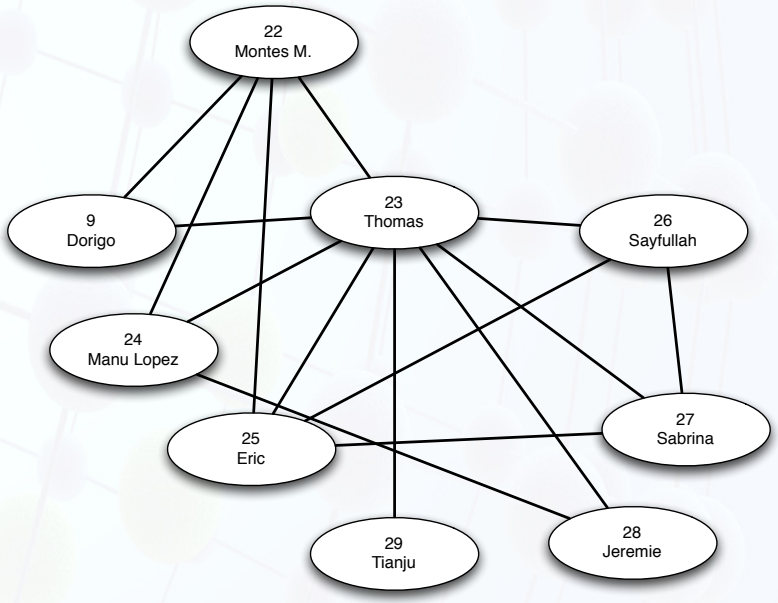




Community Detection

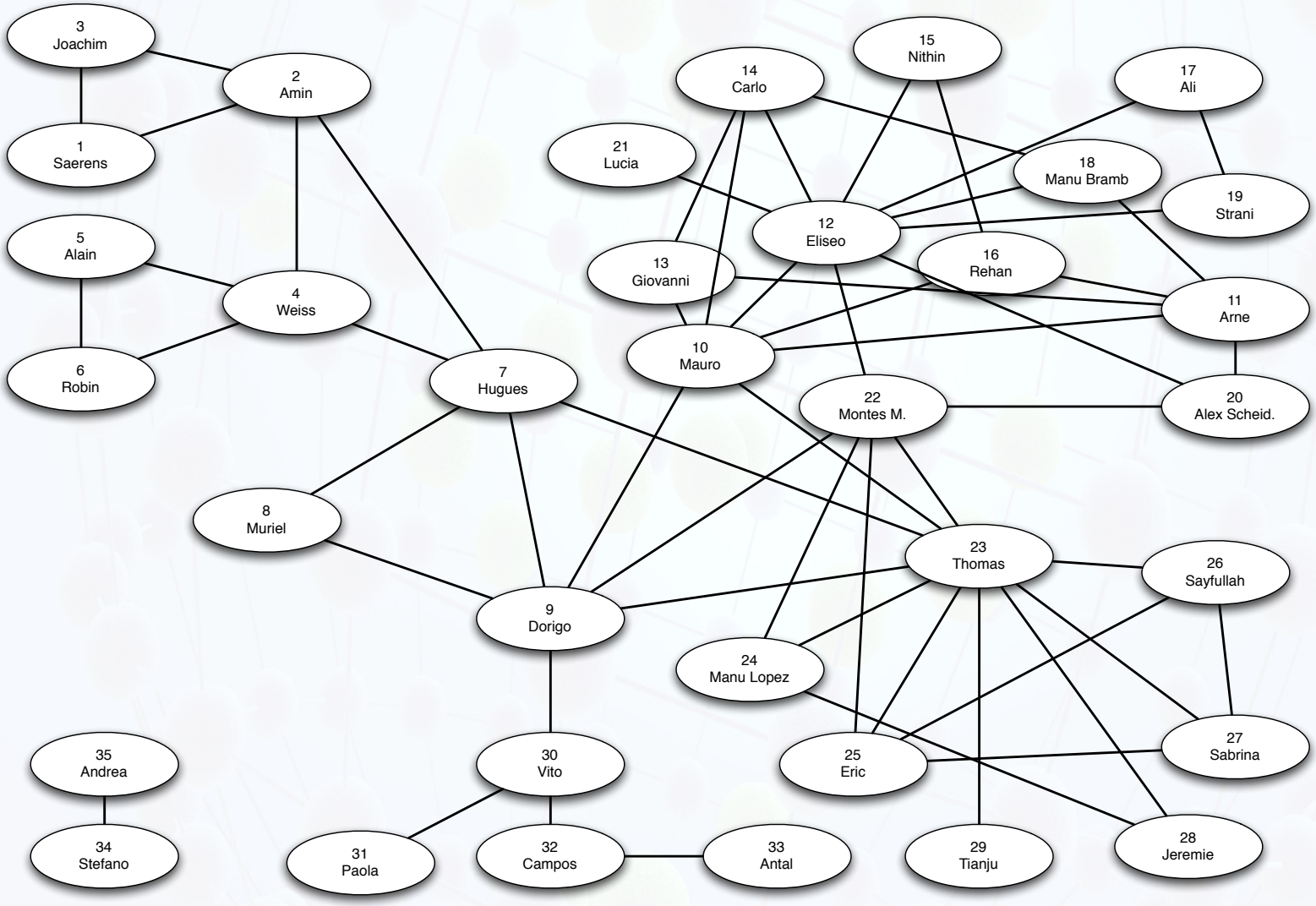


return



Application to Classification

The IRIDIA members network:



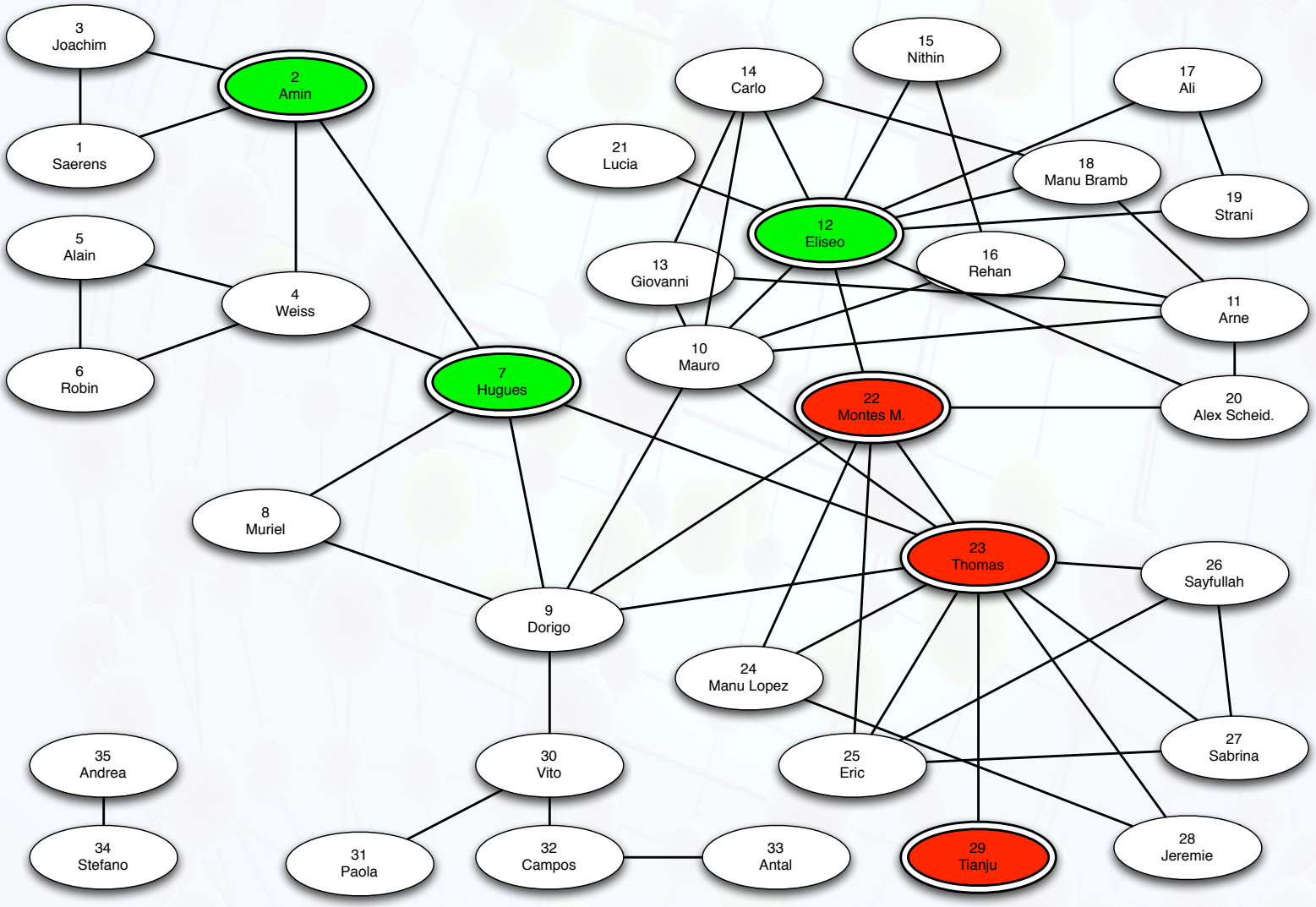
Application to Classification

After thesis reception

Who will attend to the **after thesis reception**?
This is a **within-network classification** problem.

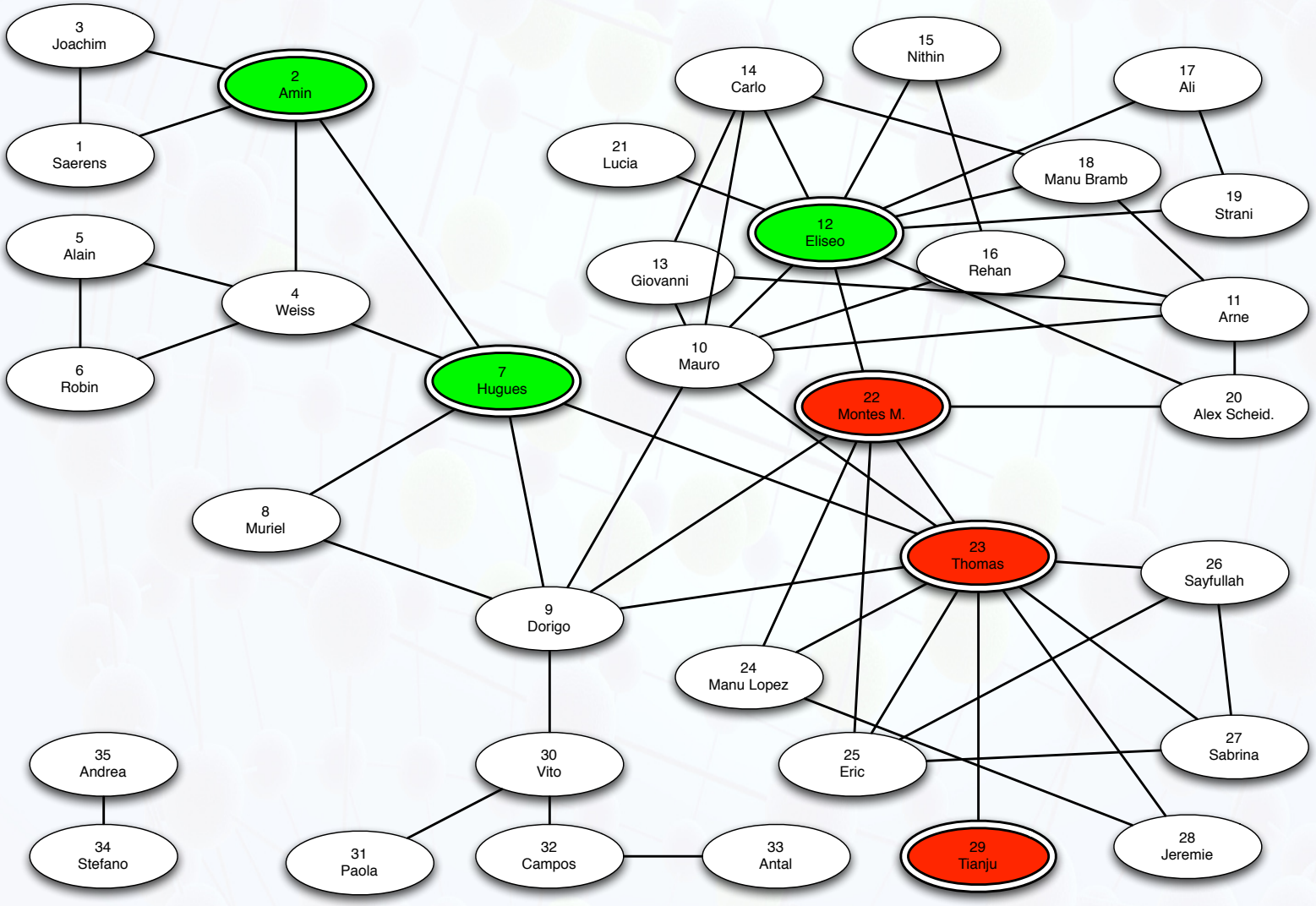
- Suppose, we know that some person will attend, and some will not attend.
- Can we predict for the others if they will attend or not?

Application to Classification



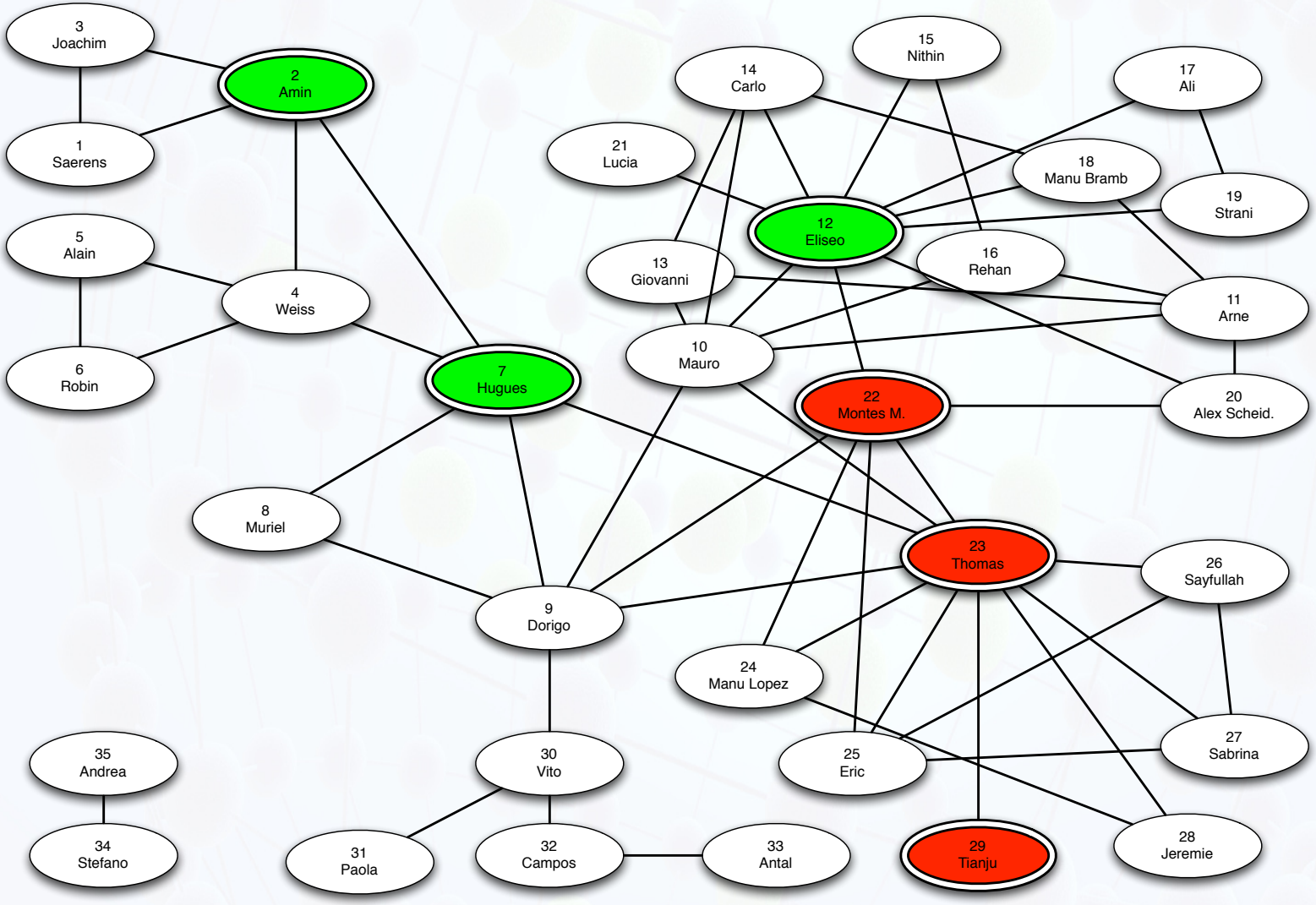
Question: Will Mauro attend to the after dinner reception?

Application to Classification



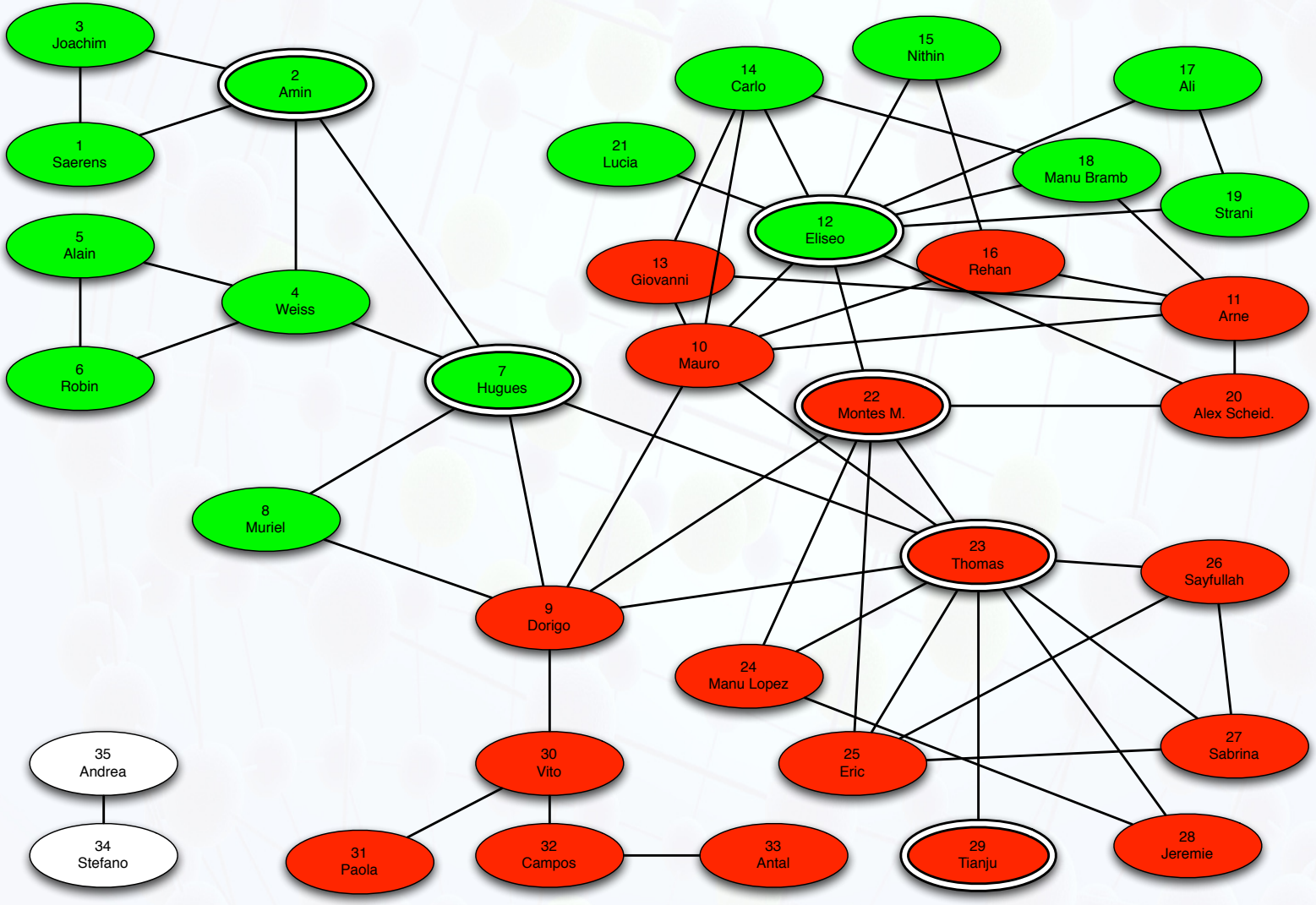
- Sum the similarity of Mauro with: Amin, Hugues and Eliseo ~ 0.74
- Sum the similarity of Mauro with: Montes, Thomas and Tianju ~ 0.97

Application to Classification



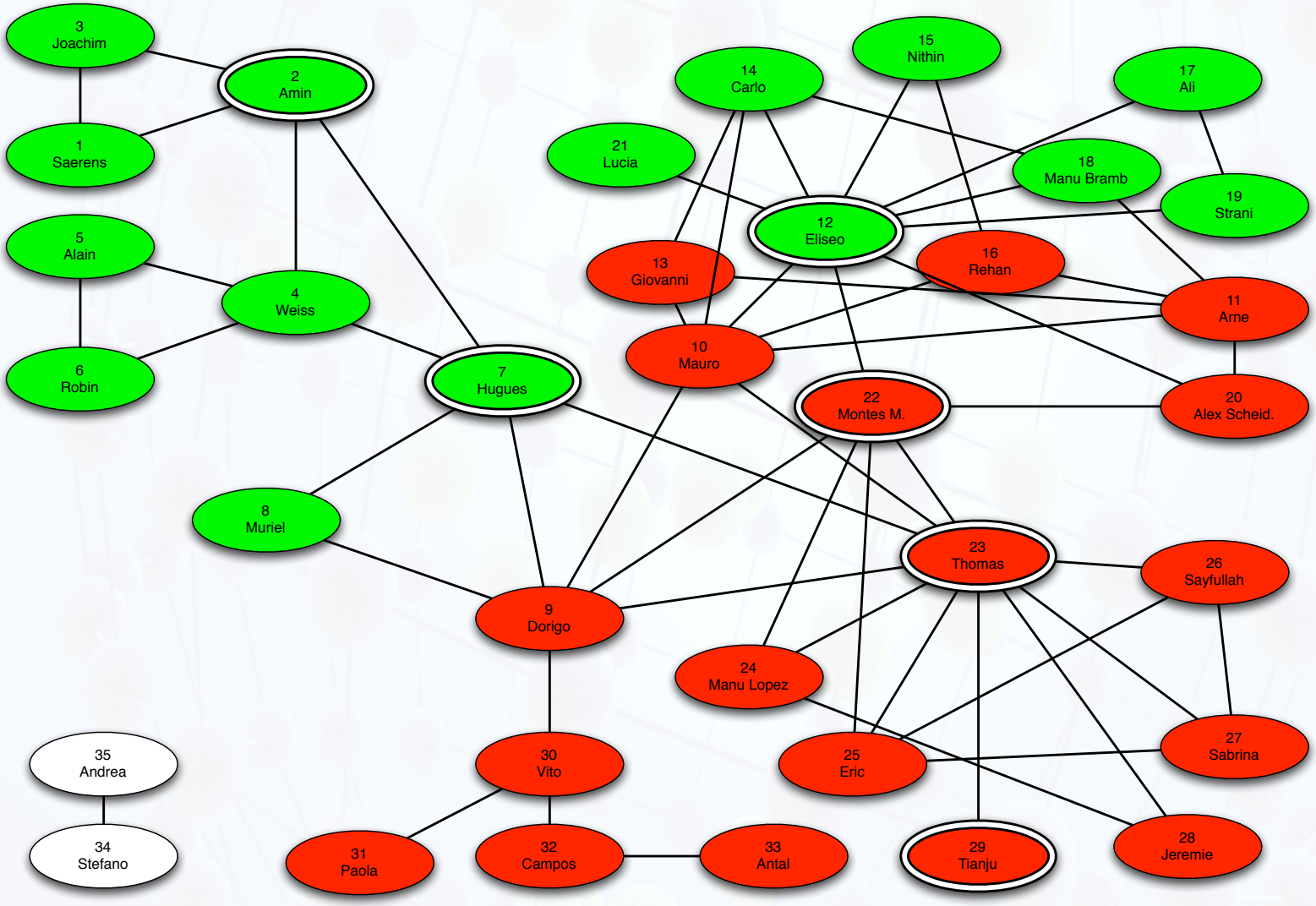
Let us classify all the nodes.

Application to Classification

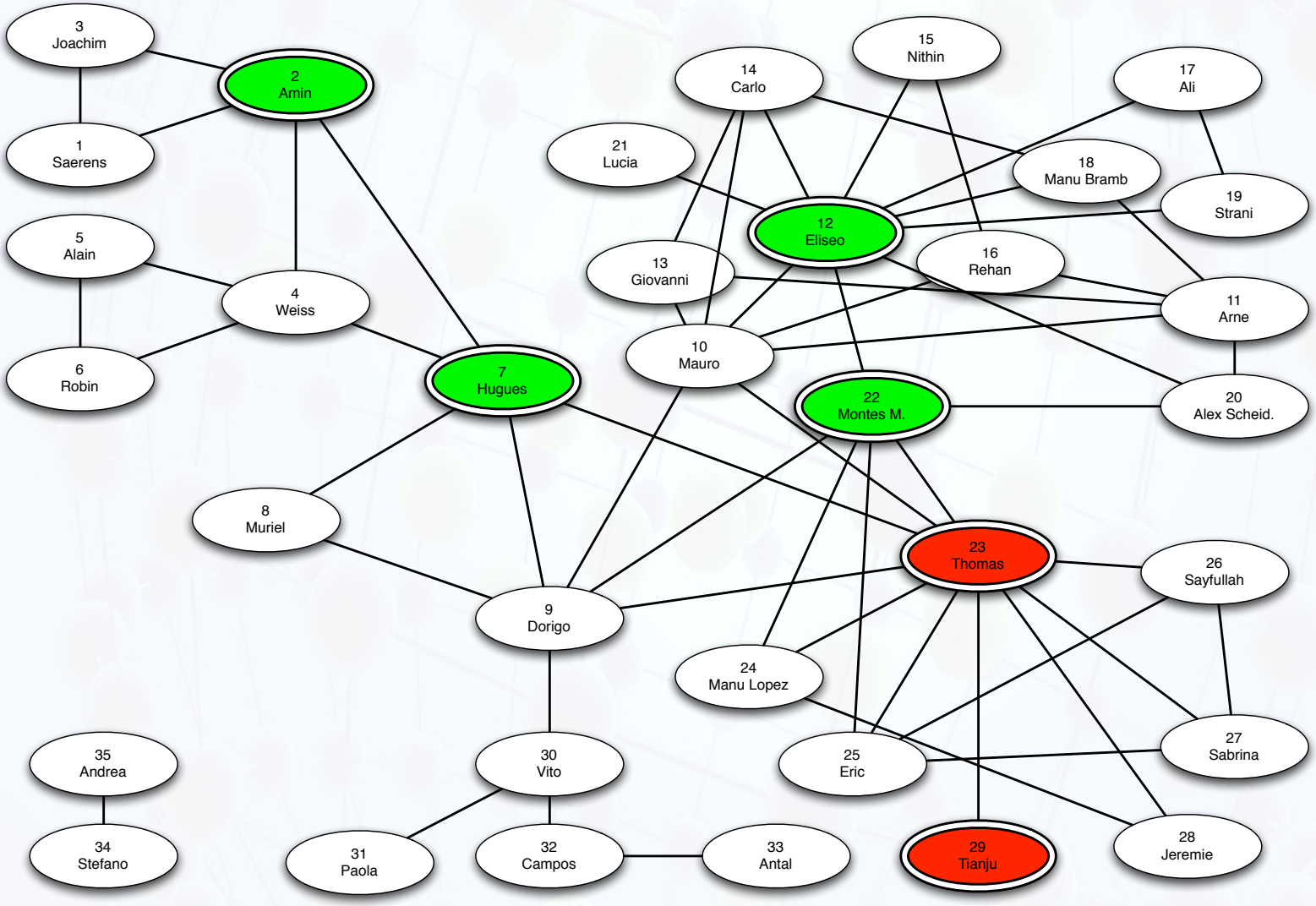


Let us classify all the nodes.

Application to Classification

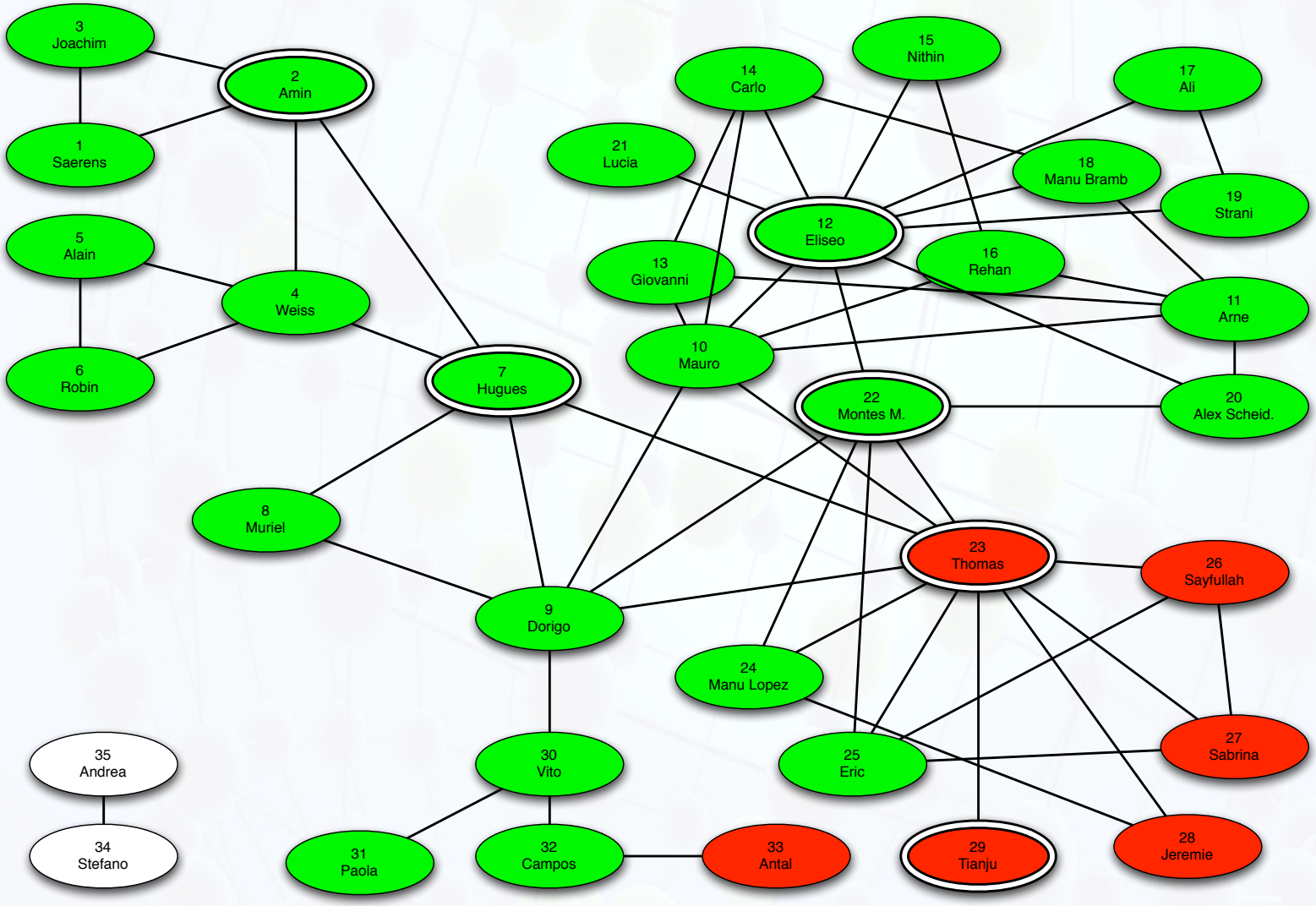


Application to Classification



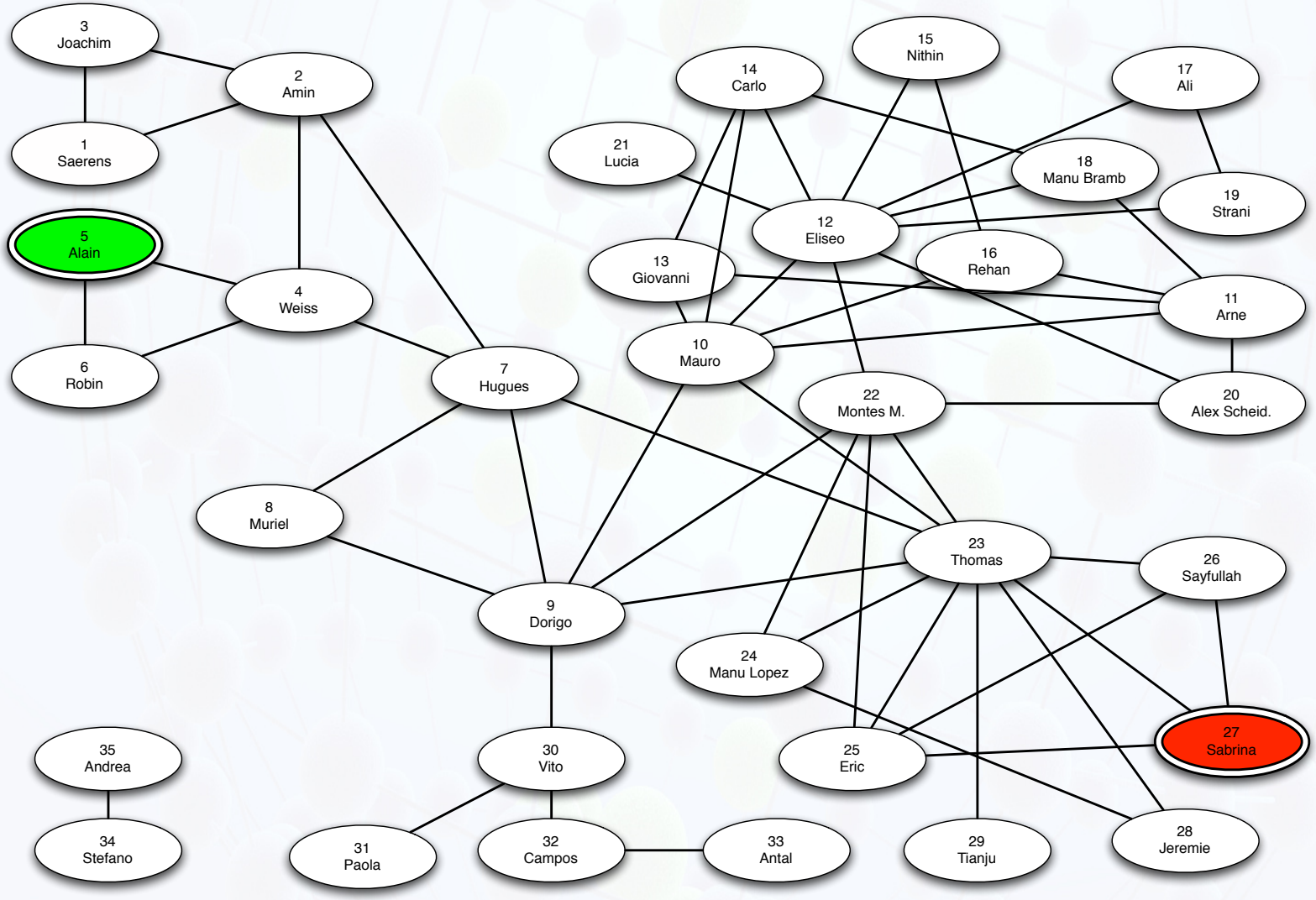
If Marco Montes attends, how does it influence the others....

Application to Classification



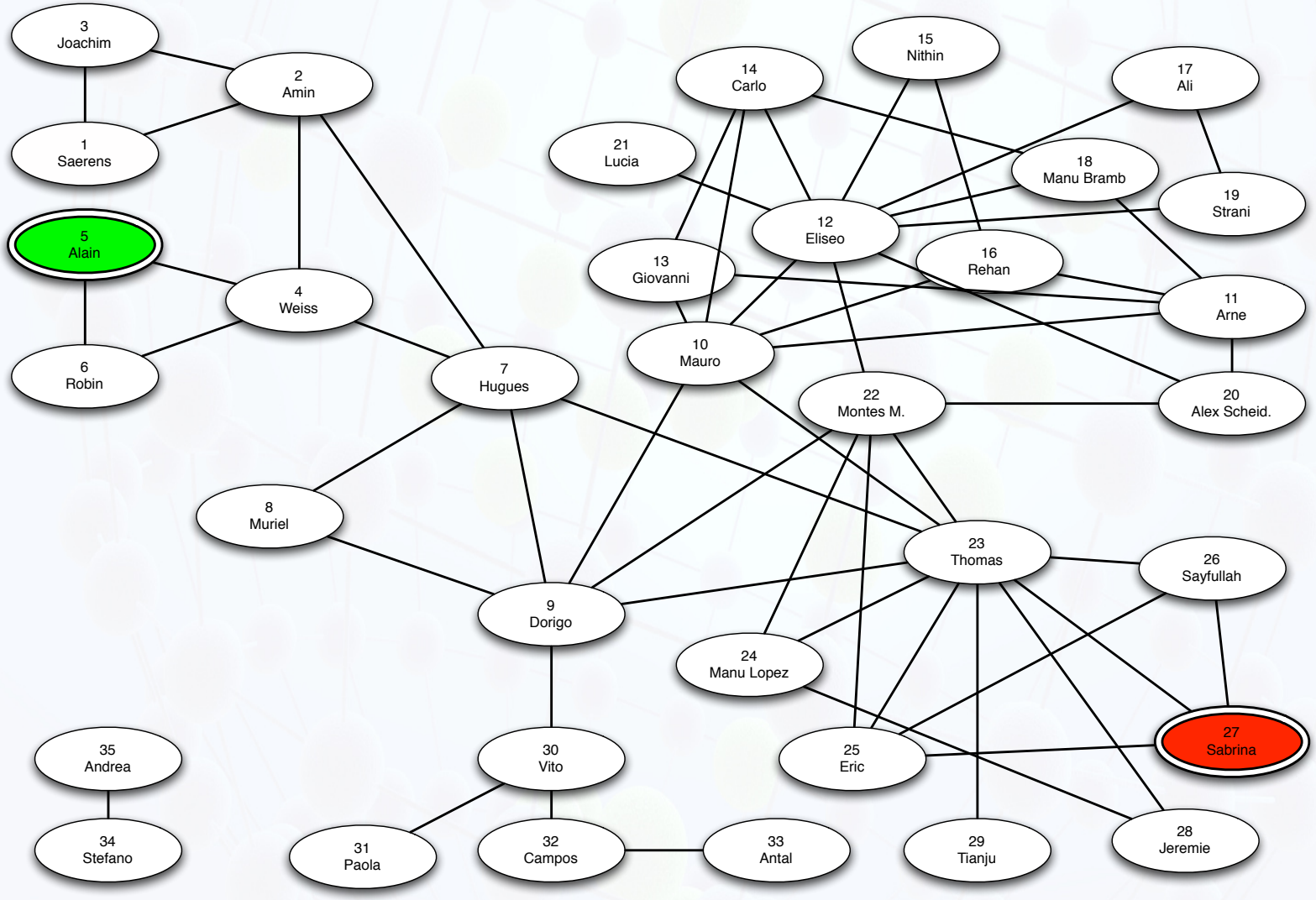
Algorithms

aSop: Two nodes are considered as highly correlated if they often appear together on the same –preferably short– path.



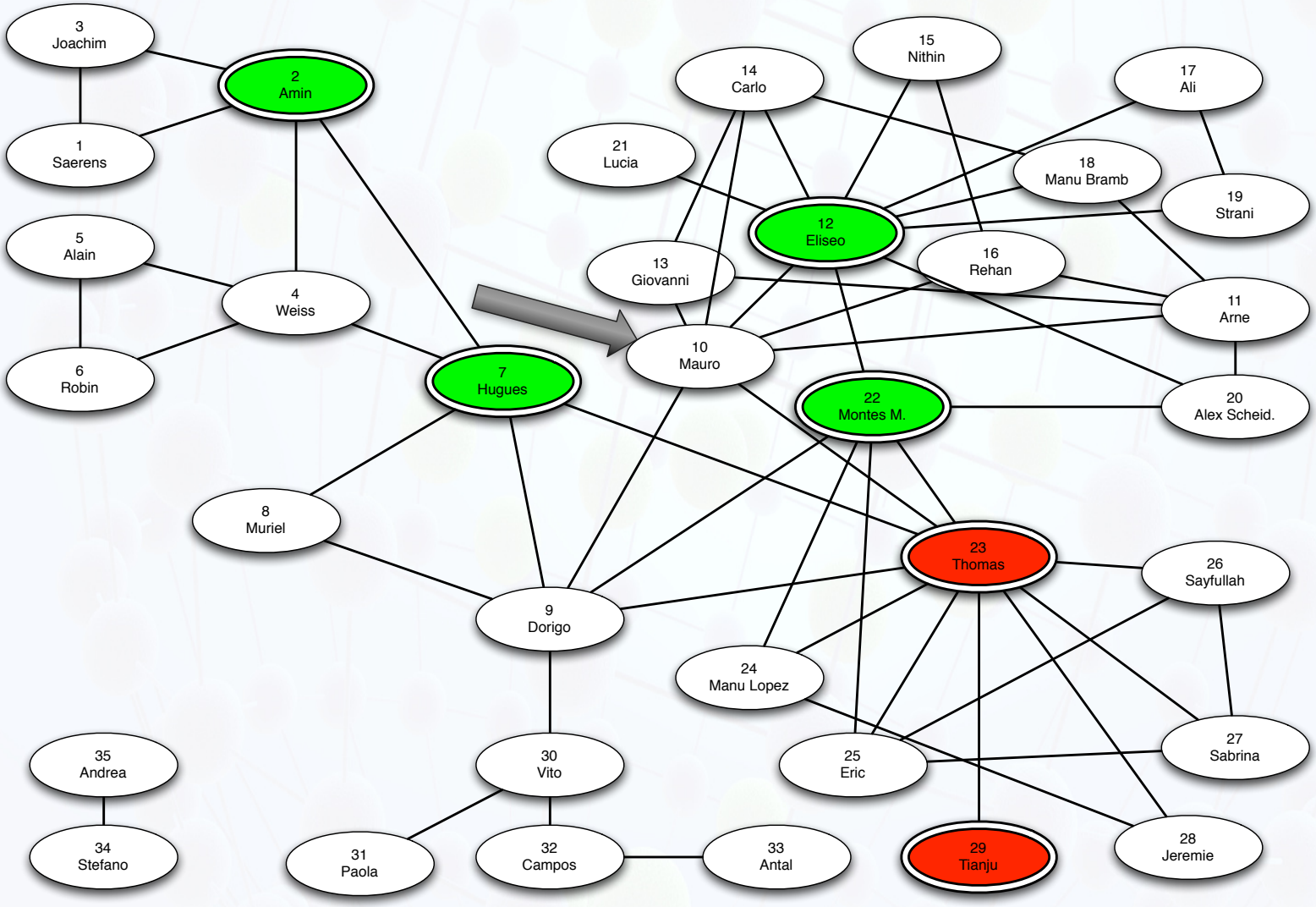
Algorithms

bNRWR: Normalized expected number of visits of node j starting from node i for walks of maximum τ steps.



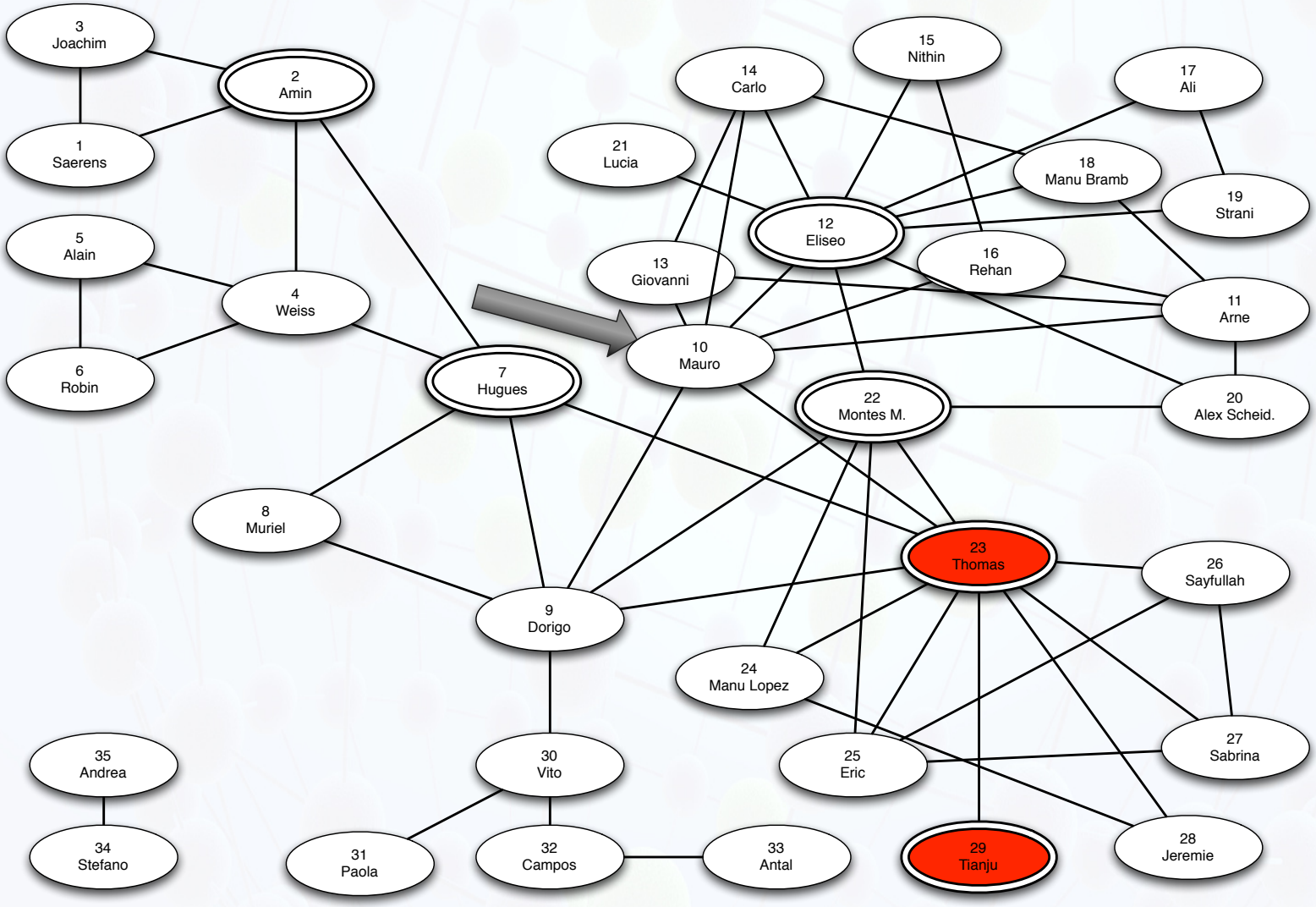
Algorithms

b$\mathcal{D}$WALK: Measures the centrality of a node inside a specific class.



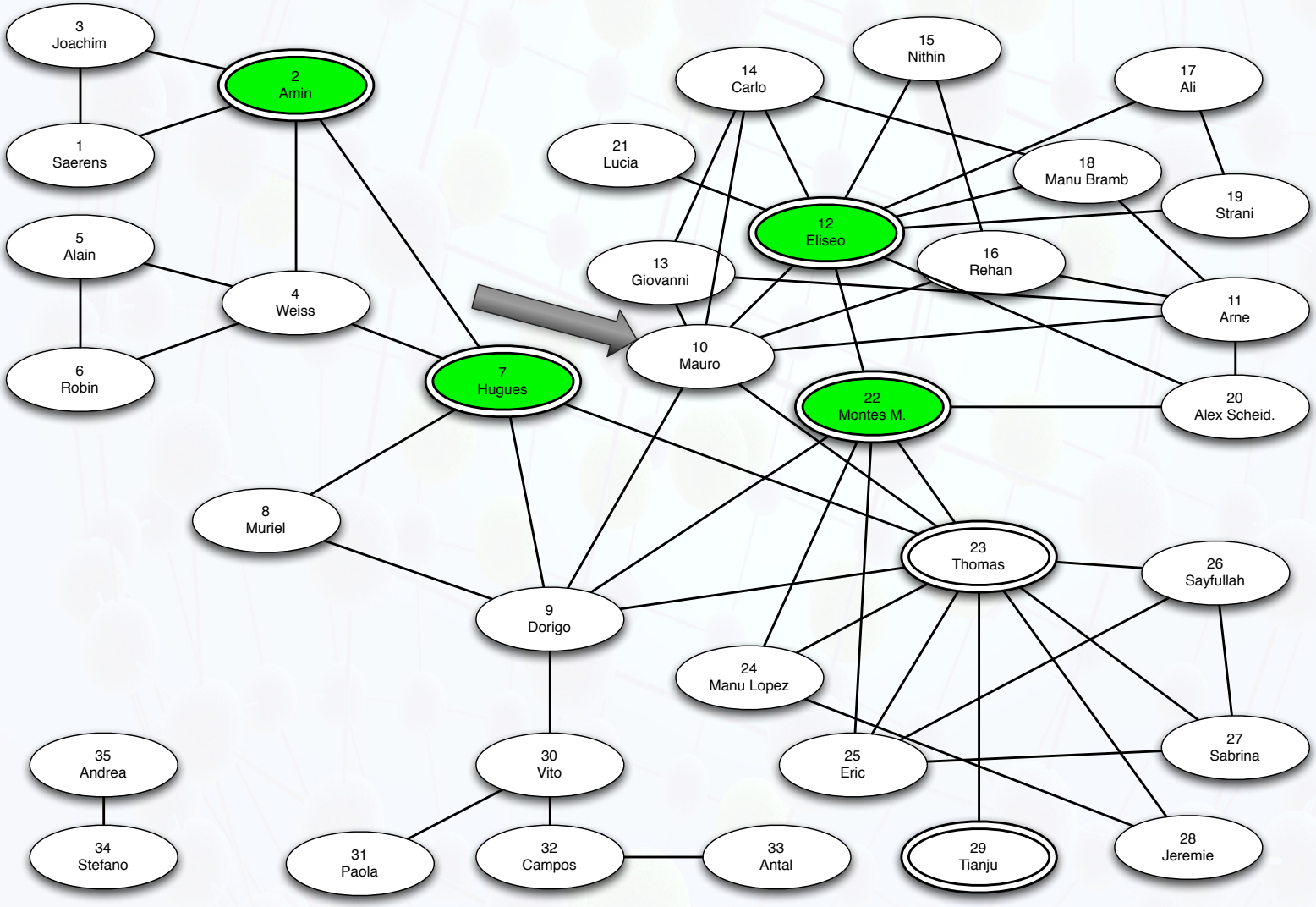
Algorithms

b$\mathcal{D}$WALK: Measures the centrality of a node inside the red class.



Algorithms

b$\mathcal{D}$WALK: Measures the centrality of a node inside the green class.



In Summary

- Time complexity: **linear** in the number of links, classes and steps $\rightsquigarrow$ **applicable on large-scale graphs**.
- Spatial complexity: **store in memory the graph and scores for each node**.

Experiment: Application on Large Network

Category	Size	Proportion
Chemicals	630107	19,42%
ICT	381537	11,76%
Drugs and medical	245595	7,57 %
Electrical and electronic	575369	17,73 %
Mechanical	724022	22,31 %
Others	688375	21,21 %
Total	3245005	100%
Majority class proportion	22,31%	

Table: **Class distribution for the `U.S. patents` data set.**

Experiment: Results

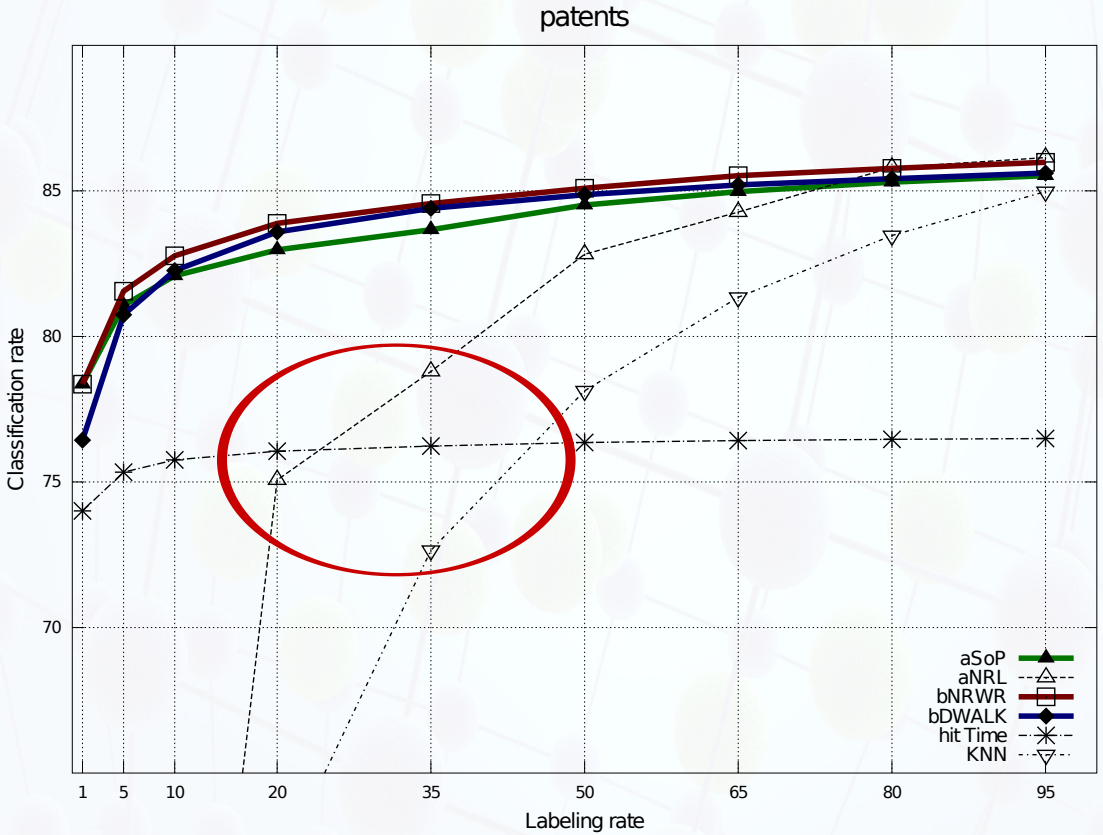


Figure: **Classification rates averaged on 5 runs for an increasing labeling rate of 10, 20, 35, 50, 65, 80 and 95%.**

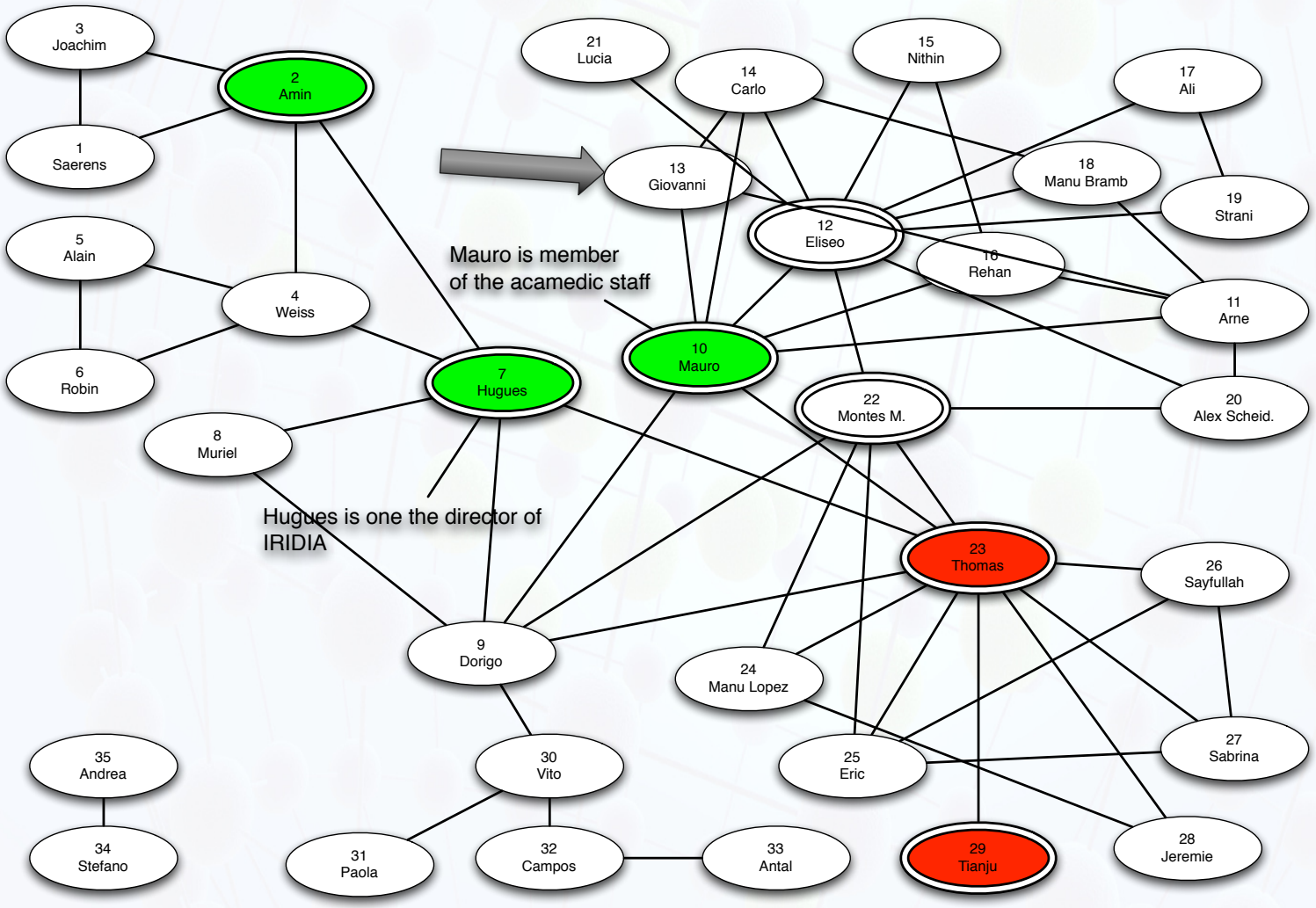
Experiment: Computation Time

Algorithm	1%	5%	10%	20%	35%	50%	65%	80%	95%
aSoP	769	749	972	883	658	291	313	351	337
aNRL	45	15	17	25	46	77	134	179	246
bNRWR	41	42	31	82	118	178	261	380	505
b$\mathcal{D}$WALK	55	58	63	79	120	184	271	379	511

Table: **Overview of cpu time in seconds needed for running an algorithm (and thus classifying all the unlabeled nodes), averaged over 10 runs, obtained on the `U.S. patents network` for labeling rates of 1, 5, 10, 20, 35, 50, 65, 80 and 95%. Results are reported for the aSoP, the bNRWR, the b$\mathcal{D}$WALK and the aNRL. The cpu used is an Intel(R) Xeon(R) CPU E5335 @2.00GHz, with 4096 KB of cache size and 8GB of RAM.**

return

Combine the Graph with the Information on Nodes



return

Global Conclusion

- We proposed a novel betweenness measure: the **SoP betweenness** which is computable in linear time on large-scale sparse directed graphs
- We proposed a novel clear and precise covariance: the **SoP covariance** which measure similarity between two nodes of a directed graph
- We introduce **three novel algorithms** for within-network classification on large-scale sparse network with a **linear complexity** in terms of labels, steps and links.
- A novel data set has been collected, the **U.S. patents**, and is now available to the community for benchmark purposes.

Perspectives

- Using the proposed measures for detecting communities in large-scale network.
- Apply graph mining techniques to **patents analysis**:
  - Detecting dense webs of patents ("patents thickets")
- Use wikipedia as external graph resource to improve classification performance

Publications...

Journal Papers

– **Amin Mantrach**, Luh Yen, Jerome Callut, Kevin Francoisse, Masashi Shimbo, and Marco Saerens. The sum-over-paths covariance kernel: A novel covariance measure between nodes of a directed graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1112-1126, June 2010

– **Amin Mantrach**, Nicolas van Zeebroeck, Pascal Francq, Masashi Shimbo, Hugues Bersini and Marco Saerens. Semi-supervised Classification and Betweenness Computation on Large, Sparse, Directed Graphs, *submitted for publication to Pattern Recognition*, PR-D-09-01097R, Minor Revision

– Caroline Herssens, **Amin Mantrach** and Marco Saerens, Ant colony optimization revisited from a randomized shortest path perspective, submitted for publication

International Conference Papers

– L. Yen, **A. Mantrach**, M. Shimbo, and M. Saerens. A family of dissimilarity measures between nodes generalizing both the shortest path and the commute-time distances. *Proceedings of the 14th SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785793, 2008.

– L. Kevers, **A. Mantrach**, C. Fairon, H. Bersini and M. Saerens (2010), Classification supervisée hybride par motifs lexicaux étendus et classificateurs SVM, *10th International Conference on statistical analysis of textual data (JADT 2010)*, Rome, 9-11/06/2010, S. Bolasco, I. Chiari, L. Giuliano ed(s), Ed. Univ. di Lettere Economia Diritto, 2010, p. 105-117.

– **Amin Mantrach** and Marco Saerens, The All-Paths Covariance: a new covariance measure between nodes of a weighted, directed, graph, *MLG 2008 - 6th International Workshop on Mining and Learning with Graphs*, ID 15.

Thank you for your attention

Questions ?