Bio-inspired Community Finding in Social Networks, new trends and challenges

DAVID CAMACHO

David.Camacho@uam.es

Computer Science Department

Autonomous University of Madrid

Applied **I**ntelligence & **D**ata **A**nalysis
http://aida.ii.uam.es
Universidad Autónoma de Madrid

Outline

- Social Network Analysis
- Community Finding Problems
- Evaluating communities
- Bio-inspired Community Detection Algorithms

Outline

- **Social Network Analysis**
- Community Finding Problems
- Evaluating communities
- Bio-inspired Community Detection Algorithms

Social Network Analysis

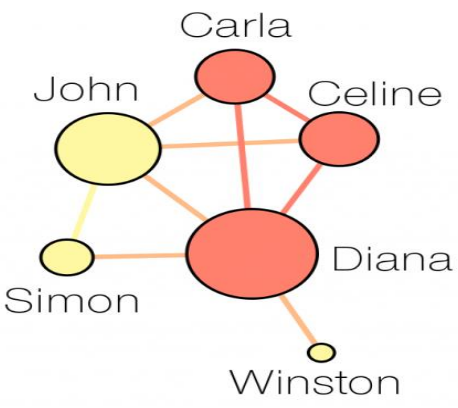
- Social Networks are created from the social **interactions of humans**
- Nowadays, SNs have gained popularity due to the famous SN platforms available on the Internet
- This popularity has generated a **huge amount** of data:
  - The users interact with each others.
  - Due to its topology and its behaviour is really simple to spread any though.
- For these reasons, companies and the research community have focused their effort to **analyse** the SNs

G. Bello-Orgaz, J. Jung, D. Camacho. **Social big data: Recent achievements and new challenges**. Information Fusion, 28: 45-59, 2016

Social Network Analysis

- **Social network analysis** (SNA) is a mathematical or computer science theory which consist of **modelling** the individuals of a network and the relationships between them, and **extracting** some valuable hidden **information** using algorithms and statistics

- Related areas: sociology, anthropology, **social psychology** (origins); **graph theory**, matrix algebra and statistics (mathematical bases); physics, **computer science**, informatics, biology (increasing)



Nodes

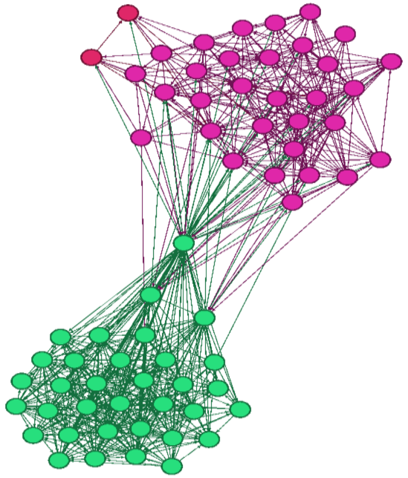
Id,Label,Gender
1,John,1
2,Carla,2
3,Simon,1
4,Celine,2
5,Winston,1
6,Diana,2

Edges

Source,Target
1,2
1,3
1,4
1,6
2,4
2,6
3,6
4,6
5,6

Social Network Analysis

- Contributions of network analysis to **online communities**
  - Identify the **group creators**
  - Identify **opinion leaders**
  - Identify **key accounts** of a network
  - Identify **information consumers** in a network
  - Identify the **main distributors of information** in the network
  - Follow the **dissemination** of a message
  - Determine the **dominant genre** of a network
  - **Multimember ship** identify an individual
  - Identify individuals keep having the **most or the relationships** in a network
  - Identify **the network of an individual (ego)**
  - **Communities detection**
  - Topic detection



Tools for SNA

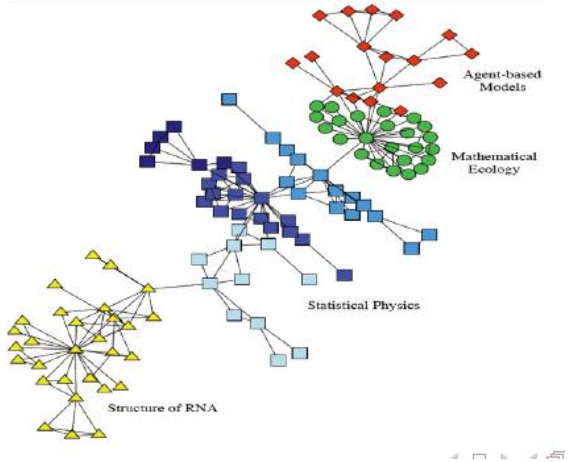
1. There exist a large number of software for SNA. Some examples are:
   - **Gephi:** https://gephi.org/
   - **Pajek**: http://mrvar.fdv.uni-lj.si/pajek/
   - **UCINET**: https://sites.google.com/site/ucinetsoftware/home
   - **KrackPlot**: http://www.andrew.cmu.edu/user/krack/krackplot.shtml
   - **GUESS**: http://graphexploration.cond.org/
   - **KDNuggets: Top 30 Social Network Analysis and Visualization Tools**
   - https://www.kdnuggets.com/2015/06/top-30-social-network-analysis-visualization-tools.html
   - **Graphviz**, **Graph-tool**, **EgoNet**, **Cuttlefish**, **InFlow**, **JUNG**, **NetMiner**, **NetworkX**, **SocNetV**, **SVAT**,...

Outline

- Social Network Analysis
- **Community Finding Problems**
- Evaluating communities
- Bio-inspired Community Detection Algorithms

Community Finding Problems

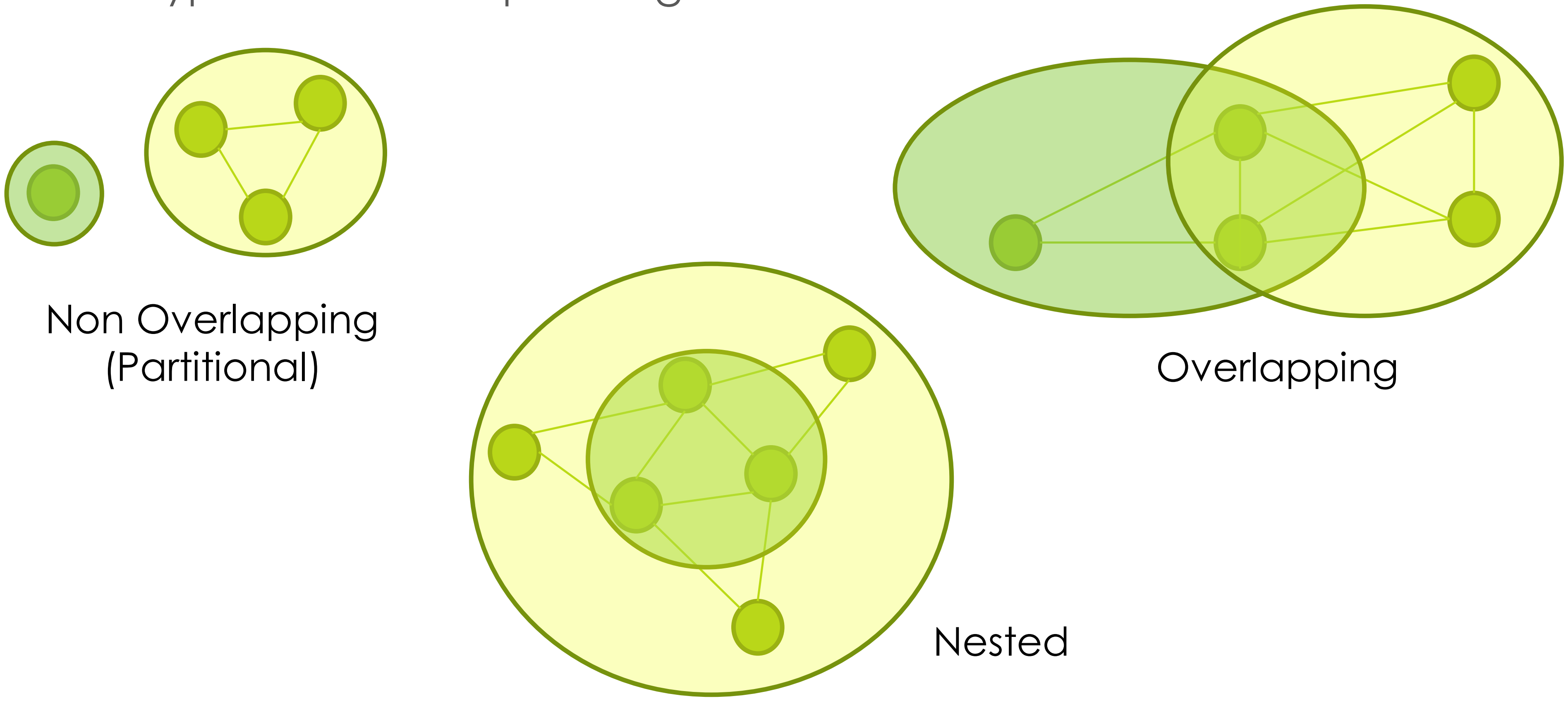
- The **Community Detection Problem** (CDP)
  - Can be defined as the **division** of a graph into **clusters** or **groups** of nodes where each one includes: a strong internal cohesion, and a week external cohesion
  - Applied in **several disciplines** such as *sociology*, *biology*, or *computer science*, whose information can be easily represented as a network or **graph**.
- Similar users belong to the same community, whereas different users are not located in the same group.
- The key question is how we define that two users are "**similar**".



Santo Fortunato. Community detection in graphs. Physics reports 486 (3-5), 75-174, 2010

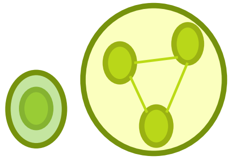
Community Finding Problems

- Types of CFPs depending on the outcome communities:

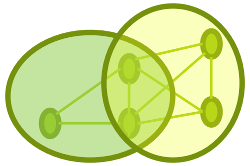


Community Finding Problems

- Types of CFPs:



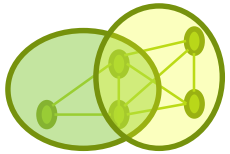
  - **Partitional methods**: a disjoint division of the graph is performed (each vertex will only belong to a single community), NP-hard problem. Popular algorithm: **Edge *Betweenness* Centrality (EBC)**, Girvan & Newman; others: **Fast Greedy** or **Louvain Method** (based on *Modularity*), **NetWalk, WalkTrap, InfoMaps** (based on Random Walks)



  - **Overlapping methods**:
    - **Fuzzy methods:** each node is associated with a community using a membership factor
    - **crisp (non-fuzzy) methods**: the relationship between a node and a cluster is binary

Community Finding Problems

- Types of CFPs:



  - **Overlapping methods crisp (non-fuzzy) methods**:
    - Clique Percolation Method (**CPM**), based on the *clique* concept
    - Cluster-Overlapping Newman Girvan Algorithm (**CONGA**) , **EBC** and **CONGO**, based on *betweenness*
    - Cluster Affiliation Model for Big Networks algorithm (**BIGCLAM**), **CoDA**, based on an statistical approach (model-based methods)

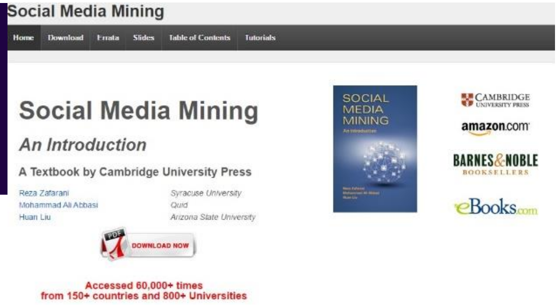
Community Finding Problems

- The key concept in CFP is to find a ‘**good**’ partition or community
- Usually the quality of any community depends on:
  - **Internal connections *(metrics)***: used ot assess/evaluate how connected are the nodes inside a community
  - **External connections (*metrics*)**: used ot assess/evaluate how disconnected are the communities
  - The goal will be finding communities with **strong** internal connections and **weak** external connections

Community Finding Problems

- **General metrics and relational properties from graph theory:**
  - Degree (in-degree, out-degree, average degree).
  - Density
  - Components, cliques and cores
  - Centrality
  - Centralisation
  - Erdos/Bacon number(s)
  - Diameter
  - …

Community Finding Problems



- **Internal connectivity (metrics from graph theory)**
  - *Triangle Partition Ration* (**TPR**): measure of graph cohesion (fraction of nodes in a graph that belongs to a triangle)

    $$TPR = \frac{|\{u : u \in V, \{(v,w) : v,w \in V, (u,v) \in E, (u,w) \in E, (v,w) \in E\} \neq \emptyset\}|}{|V|}$$

  - *Local Clustering Coefficient* (**LCC**): measures the transitivity of a node into the graph

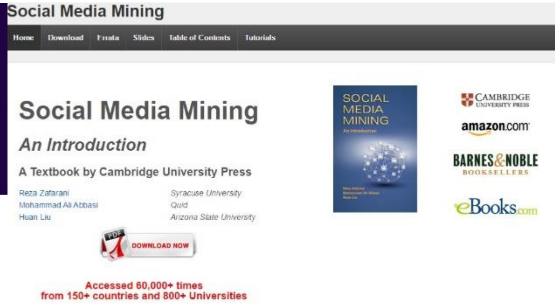
    $$LCC_i = \frac{2 \times \sum_{j,h} a_{jh} a_{ij} a_{ih} a_{ji} a_{hi}}{k_i(k_i - 1)}$$

  - *Global Clustering Coefficient* (**GCC**): measures the global transitivity of the graph $$GCC = \frac{3 \times |Triangles|}{|Triples|}$$

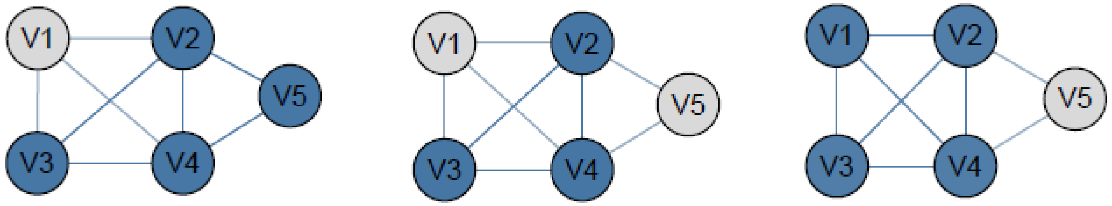
  - *Density* (**D**): measures how dense (n° edges/nodes) is the graph

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

Community Finding Problems



- **Internal connectivity (metrics from graph theory)**
  - *Clique Number* (**CN**): A *clique* of a graph is a subset of mutually adjacent vertices in *V* (every two vertices in the subset are connected by an edge). A clique is **maximal** if it is not contained by any other clique.

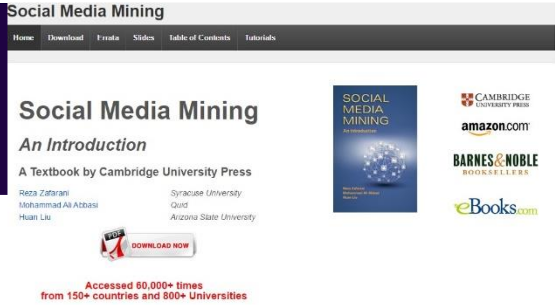


not a clique — 3-clique (not maximal) — 4-clique (maximal)

  - *Heterogeneity* (**H**): $H = \frac{\sqrt{variance(k)}}{mean(k)}$
  - *Centralization* (**Cd**): measures how **central** is a node (based on centrality metric)

$$C_d(G) = \frac{\sum_{i=1}^{n} [C_d(v*) - C_d(v_i)]}{n^2 - 3n + 2}$$

Community Finding Problems



- **External connectivity (from graph theory)**
  - *Expansion* (**Exp**): **average** number of **external edges per node**

$$Exp(C) = \frac{|\{(u,v) \in E : u \in C, v \notin C\}|}{|C|}$$

  - *Separability* (**Sep**): **ratio** between **internal** and **external** nodes

$$Sep(C) = \frac{|\{(u,v) \in E : u \in C, v \in C\}|}{|\{(u,v) \in E : u \in C, v \notin C\}|}$$

  - *Cut Ratio* (**CR**): **ratio** between **external** edges and **all the possible external edges**

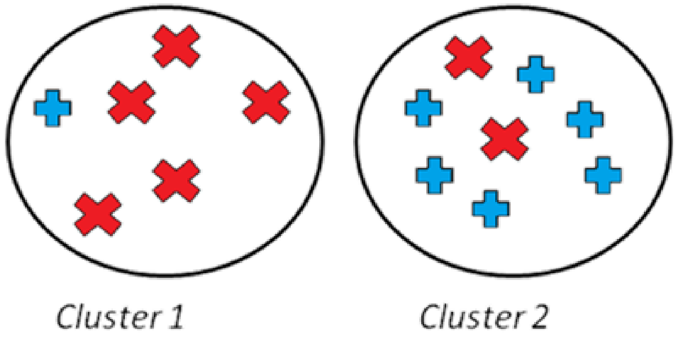
$$CR(C) = \frac{|\{(u,v) \in E : u \in C, v \notin C\}|}{n_c(n - n_c)}$$

Outline

- Social Network Analysis
- Community Finding Problems
- **Evaluating communities**
- Bio-inspired Community Detection Algorithms

Evaluating communities

- Once the CFA generates a set of communities it's essential to evaluate their **quality**.
- This is a classical problem in **Graph clustering**



- There exist two possible situations:
  - Evaluation **with ground truth**
  - Evaluation **without ground truth**

Evaluating communities

- When **ground truth** is **available**, we have at least partial knowledge of what communities should look like.
  - We are given the correct community (clustering) assignments.

- Measures:
  - Average **F1 Metric**
  - **Omega Index** (overlapping version of the Adjusted Rand Index (ARI); Hubert and Arabie 1985)

$$\omega_e(C_1, C_2) = \frac{1}{M^2} \sum_{j=0}^{max(K_1,K_2)} |t_j(C_1)| \cdot |t_j(C_2)|$$

  - **Normalized Mutual Information** (**NMI**): is a metric related to the Information Theory that can be used to calculate the **similarity between two graph partitions**

$$NMI(X|Y) = 1 - [H(X|Y) + H(Y|X)]/2.$$

  - **Accuracy** in the number of communities

JIERUI XIE et. al. **Overlapping Community Detection in Networks: the State of the Art and Comparative Study**. 2012

Evaluating communities

- When the **ground truth** is **unknown**, there are lots of different metrics to evaluate the solutions.
  - **Modularity (Q)** is the most used and best known *quality measure* for graph clustering techniques. This measure is based on the idea that a *random graph is not expected to have a cluster structure*.
    - **Q** it's widely used in CFA, but calculate Q is a Np-hard problem

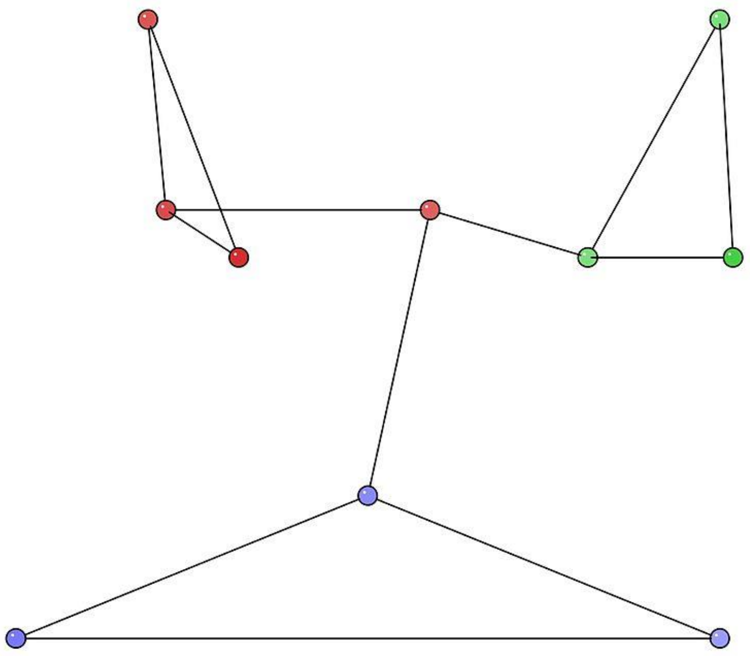
$$Q = \frac{1}{(2m)} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{(2m)} \right] \delta(c_v, c_w) = \sum_{i=1}^{c} (e_{ii} - a_i^2)$$

  - Scoring functions based on *internal connectivity*: **internal density**
  - Scoring functions based on *external connectivity*: **expansion**
  - Scoring functions that *combine internal and external connectivity*
  - Scoring function based on a *network model*

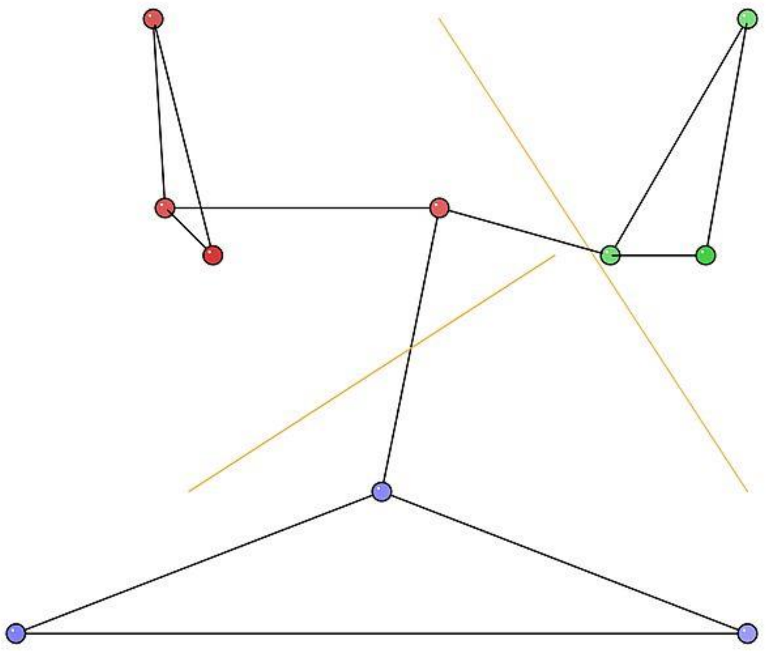
JIERUI XIE et. al. **Overlapping Community Detection in Networks: the State of the Art and Comparative Study**. 2012

Evaluating communities

- **Modularity (Q)**



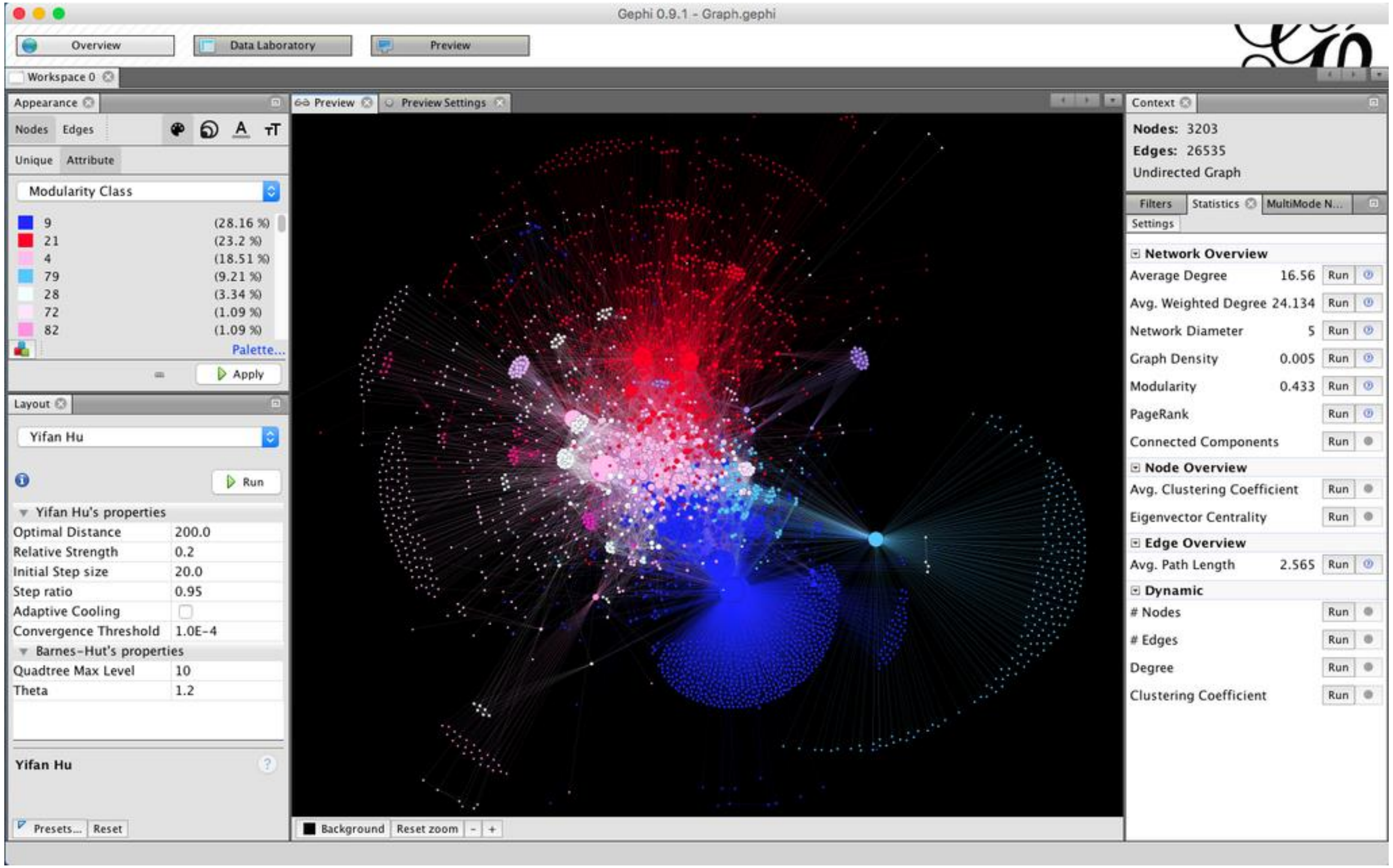
Sample Network corresponding to the Adjacency matrix with 10 nodes, 12 edges



Network partitions that maximize Q. Maximum Q=0.4896

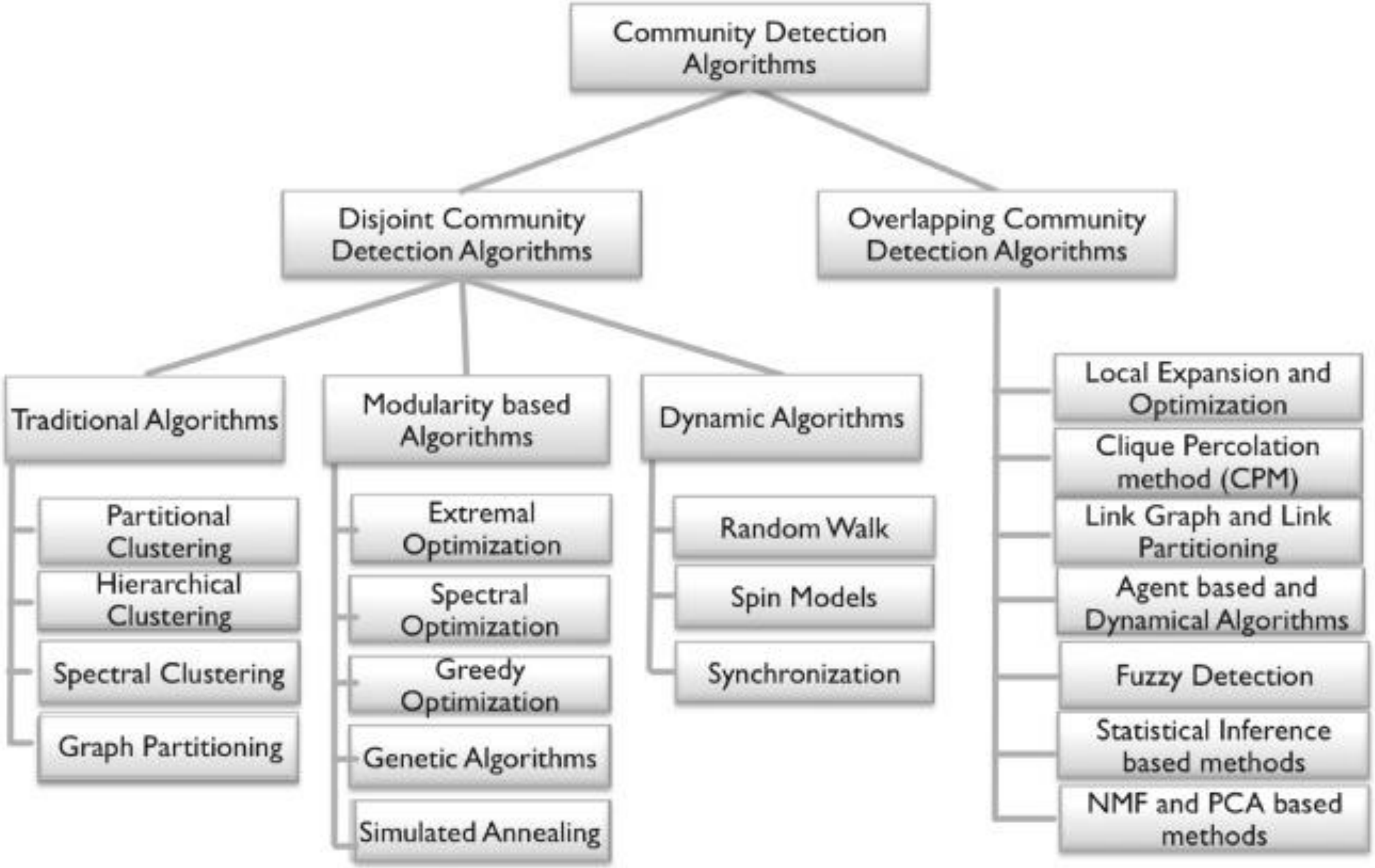
Evaluating communities

- **Modularity (Q)**



Example of modularity measurement and colouring on a scale-free network

Community Detection Algorithms



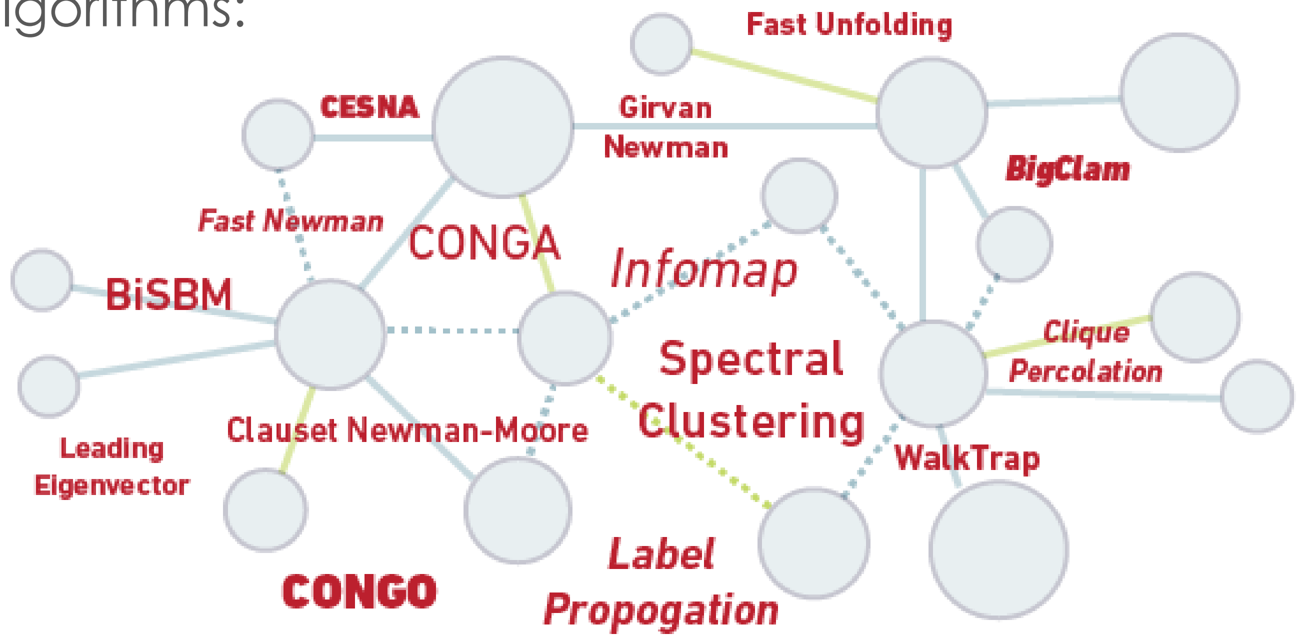
Javed et.al. **Community detection in networks: A multidisciplinary review**. Journal of Network and Computer Applications. Vol. 108, 2018

Tools to perform Community Finding Problems

- **iGraph**: http://igraph.org (7 community detection algorithms: Edgebetweenness, Walktrap Leading Eigenvectors, Fast Greedy, Label Propagation, Louvain, Spinglass, InfoMap. R, C/C++ and Python versions)
- **Gephi**: https://gephi.org/
- **JUNG**: http://jung.sourceforge.net
- **SNAP** library (Stanford Network Analysis Platform): http://snap.stanford.edu (includes Modularity, Girvan-Newman and Clauset-Newman-Moore algorithms)

Tools to perform Community Finding Problems

- **CIRCULO**: https://www.lab41.org/circulo-a-community-detection-evaluation-framework/
- A large set of algorithms:

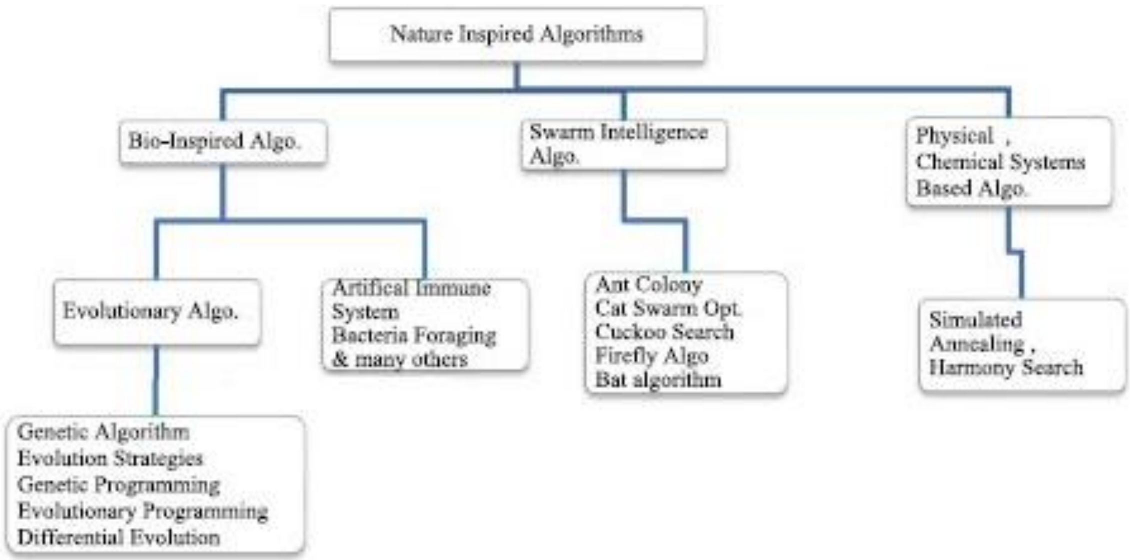


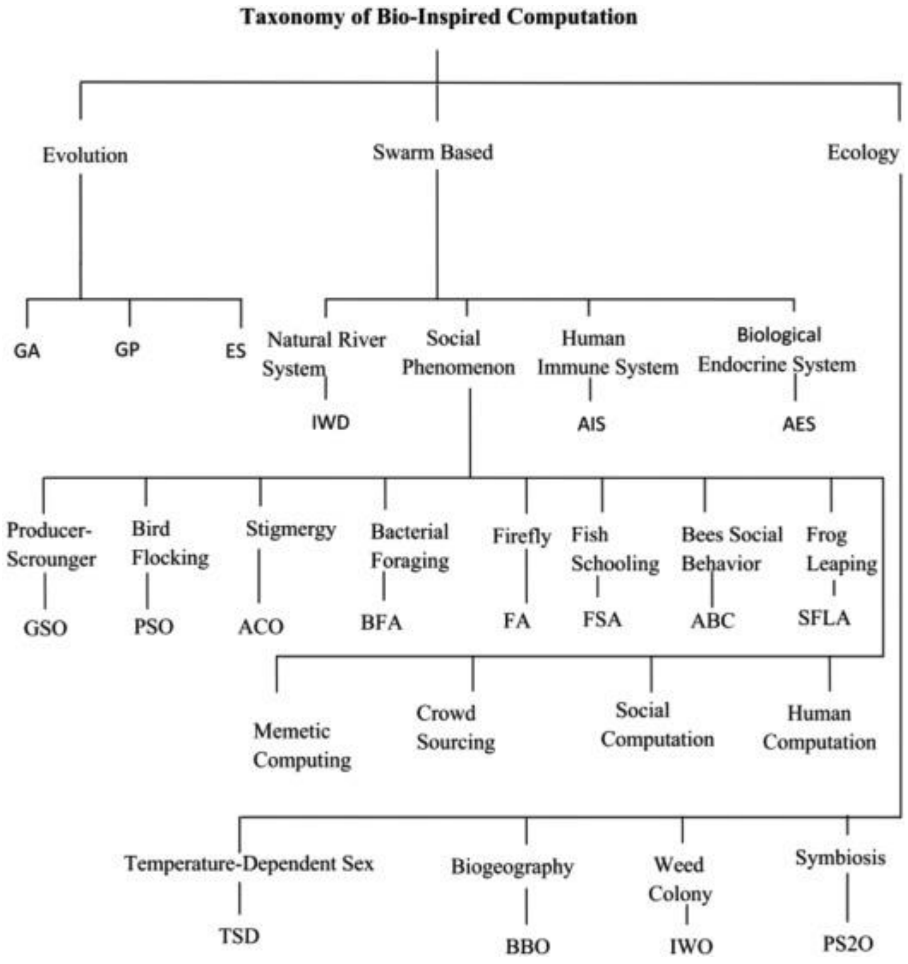
Outline

- Social Network Analysis
- Community Finding Problems
- Evaluating communities
- **Bio-inspired Community Detection Algorithms**

Bio-inspired CDAs

- We have a plenty of nature-based and bio-inspired algorithms…

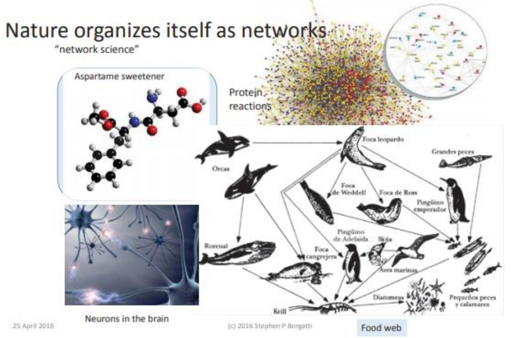


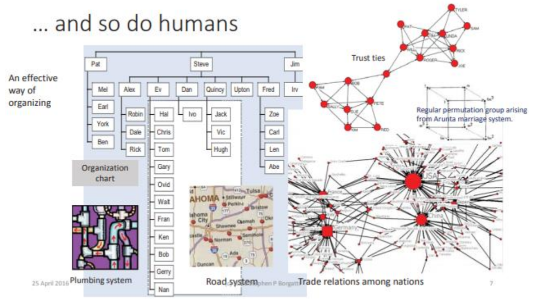


Evolutionary Computation Bestiary
https://github.com/fcampelo/EC-bestiary

Bio-inspired CDAs

- **Why is interesting to apply Nature and Bio-inspired algorithms to SNA?**
  - Most of SNA-based problems (as **CDP**) are usually **NP-hard**
  - Classical (graph mining or social mining) algorithms ***do not work well*** with Large or Very Large Networks
  - In some problems a **sub-optimal solution** would be enough
  - Some SNA-based problems needs from:
    - **Heuristics**
    - **Knowledge from the context** (not only the network structure!)
    - Work with several **criteria or objectives**
  - Bio-inspired algorithms are particularly good to manage previous problems





Bio-inspired CDAs

- **Most of them have been used in SNA and CDPs…**
  1. "A ***Multi-Objective Genetic Algorithm*** *for overlapping community detection based on edge encoding*". Bello-Orgaz, Sancho Salcedo, David Camacho. **Information Sciences**. Vol. 462, pp. 290-314, 2018
  2. "***ACO****-based clustering for Ego Network analysis*". Antonio Gonzalez-Pardo, Jason J. Jung, David Camacho. **Future Generation Computer Systems**, Vol. 66, pp. 160-170, 2017.
  3. "*Adaptive k-means algorithm for* ***overlapped graph clustering***". G Bello-Orgaz, HD Menéndez, D Camacho. **International journal of neural systems.** Vol. 22 (05), 1250018, 2012
  4. "***Firefly*** *algorithms for multimodal optimization*". Xin-She Yang. **International symposium on stochastic algorithms**, 2009. p. 169-178.
  5. "*Small worlds and mega-minds: effects of neighborhood topology on* ***particle swarm performance***". James Kennedy. **IEEE conference on Evolutionary Computation**, 1999. p. 1931-1938.
  6. "*Community detection in social networks with* ***genetic algorithms***". Clara Pizzuti. **Proceedings of the 10th annual conference on Genetic and evolutionary computation**. ACM, 2008. p. 1137-1138.

Bio-inspired CDAs

- Our work on CDAs based on GA/MOGAs
  - **GA (mono objective)**:
    - GCF-I (based on distance)
    - GCF-II (based on network topology metrics)

"Adaptive k-means algorithm for overlapped graph clustering". G Bello-Orgaz, HD Menéndez, D Camacho. ***International journal of neural systems****. Vol. 22 (05), 1250018, 2012*

  - **MOGA (multi objective, overlapping)**:
    - Node-based MOGA-OCD
    - Edge-based MOGA-OCD

"A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding". Gema Bello-Orgaz, Sancho Salcedo, David Camacho. ***Information Sciences****. Vol. 462, pp. 290-314, 2018*

Bio-inspired CDAs

- Genetic Graph-based Approaches (K-fixed GCF-I)

K-fixed GCF-I

- Based on a **standard GA**.
- Vectorial **Binary** encoding.
- **K** parameter is **fixed** as input of the algorithm.

Encoding

- **Each chromosome** is used to represent a **community**.
- **Each allele** represents the membership of a **node** into the graph.
- The **chromosome length** will be equal to the number of nodes belonging to the graph.

Gema Bello-Orgaz, HD Menéndez, D Camacho.“**Adaptive k-means algorithm for overlapped graph clustering**”. International journal of neural systems. Vol. 22 (05), 1250018, 2012

Bio-inspired CDAs

- Our method to Multi-Objective Genetic Approaches for OCD:

  - **Two new MOGA approaches** whose main **difference** lies in the **encoding**:
    1. **Node-based MOGA-OCD** the encoding represents the nodes of the graph.
    2. **Edge-based MOGA-OCD** where a new encoding schema based on the edges is used.
  - **Optimizes two** different **objective** functions:
    1. To **maximize** the **internal connectivity**.
    2. To **minimize** the **external connections** to the rest of the graph.
  - **A comparative assessment** of several **connectivity metrics** has been carried to select the most appropriate.
  - Finally, the two new algorithms have been **evaluated against** other **well-known algorithms from the state of the art.**

Bio-inspired CDAs

Algorithm 2: Multi-Objective Genetic Algorithm for Overlapping Community Detection (MOGA-OCD).

Input: A graph $N = (V, E)$ where V represents the set of vertices denoted by $\{v_1, \ldots, v_n\}$, and E is the set of edges E denoted by e_{ij} that represents a connection between the vertices v_i and v_j. Parameters m_{int} and m_{ext} represent the connectivity metrics used as the optimization criteria. Positive numbers *ngen*, *nconv* μ, λ, *mutpb* and *indmutpb* represents the main MOGA parameters to be fixed.

Output: PoF contains the best individuals

```
C ⟵ Ø;
i ← 0;
convergence ← 0;
while i ≤ ngen ∧ convergence = 0 do
    if ngen = 0 then
        C ← InitRamdomPop(λ);
    else
        C ← Cbest;
        for j ← μ to λ do
            ind1 ← RandomSel(Cbest);
            ind2 ← RandomSel(Cbest);
            ind1, ind2 ← Crossover(ind1, ind2);
            mutchoice = Random(0, 1);
            if mutchoice < mutpb then
                ind1 ← Mutation(ind1, indmutpb);
            mutchoice = Random(0, 1);
            if mutchoice < mutpb then
                ind2 ← Mutation(ind2, indmutpb);
            C ← C ∪ {ind1, ind2};
    F ← Fitness(C, m_int, m_ext);
    Cbest ← NonDominatedSortAndSelNBestNSGA2(C, F, μ);
    i ← i + 1;
    convergence ← CheckConvergence(Cbest, nconv);
return ParetoFront(C);
```

Node-based: a new node is randomly selected from the graph
Edge-based: a new edge is randomly selected from the set of the *adjacent edges*

Gema Bello-Orgaz, Sancho Salcedo, David Camacho. "**A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding**". **Information Sciences**. Vol. 462, pp. 290-314, 2018

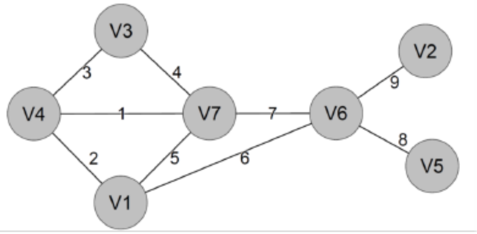
Bio-inspired CDAs

- Multi-Objective Genetic Approaches (**Edge-based**): **encoding**

Real world networks tend to be **sparse** and the **node-based methods** often have **difficulties** to find **large communities**.

Edge-based Encoding

- **Vectors** of **integer** values whose **size** are equal to the **total number of edges**.
- The **position of each allele** correspond to a **edge**.
- **Each allele** takes a **value** between the **adjacent edges** to the specified edge of this position.

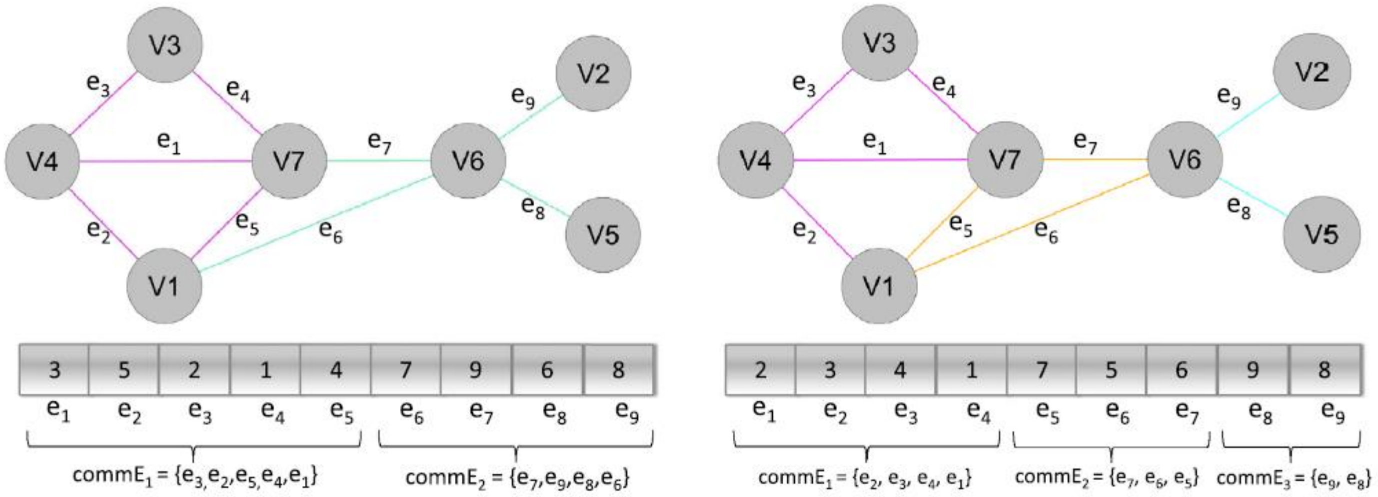


3	5	2	1	4	7	9	6	8
1	2	3	4	5	6	7	8	9

Gema Bello-Orgaz, Sancho Salcedo, David Camacho. "**A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding**". **Information Sciences**. Vol. 462, pp. 290-314, 2018

Bio-inspired CDAs

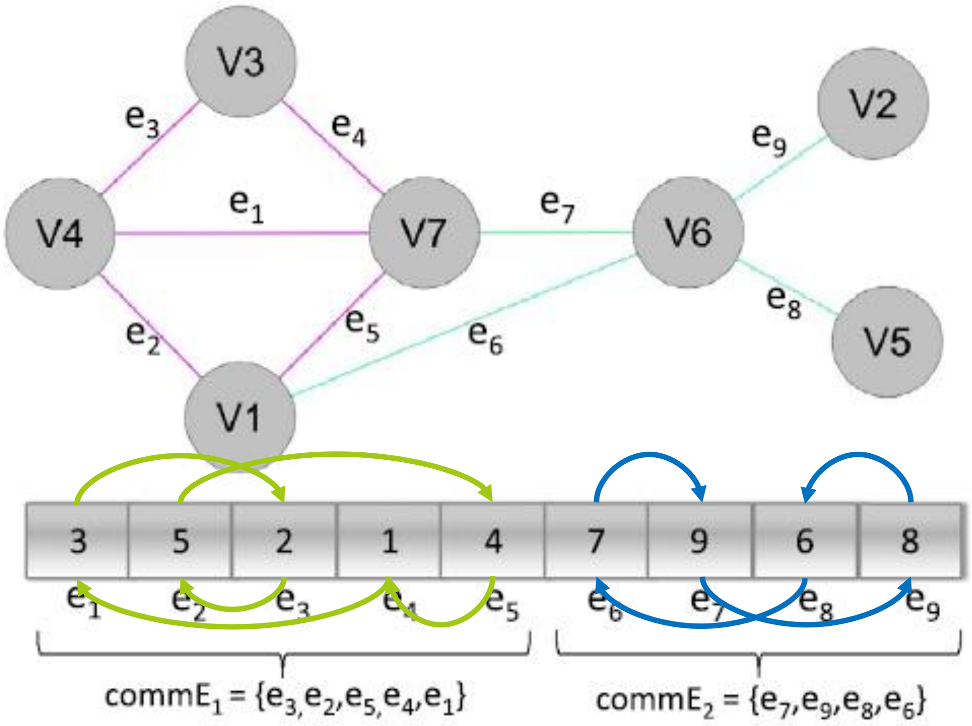
- Multi-Objective Genetic Approaches (**Edge-based**)
  1. There exist **multiple representations** for the same individual
  2. Each allele represents a <u>*specific edge of the graph*</u>, and the value of **K** (number of communities to find in the graph) has been directly encoded as part of the chromosome
  3. A **decoding phase** (from communities of edges ➡ communities of nodes) is needed

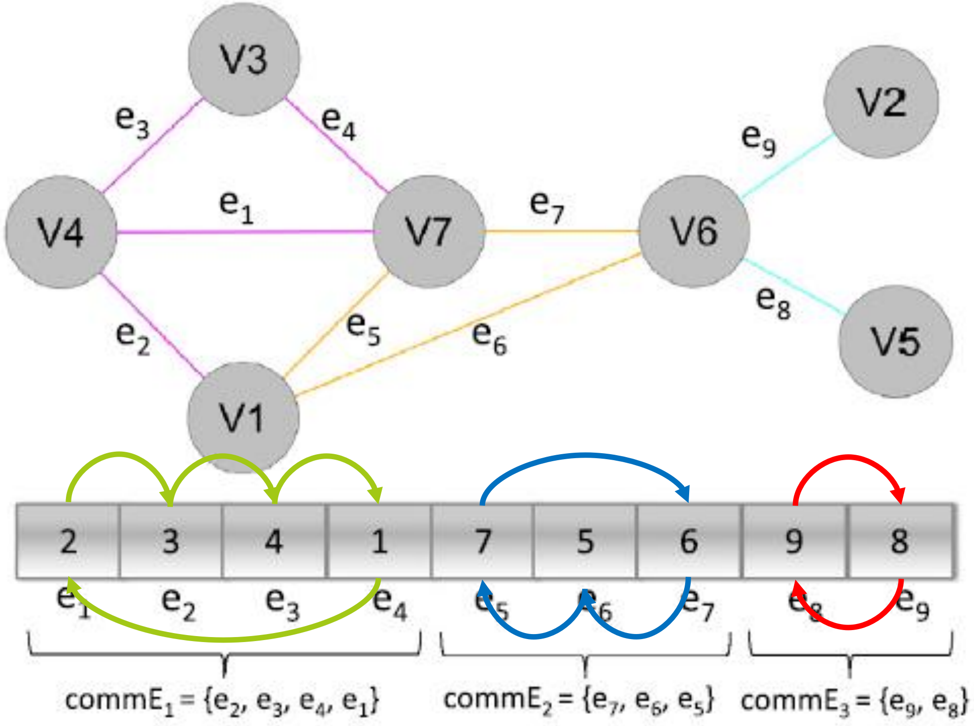


Gema Bello-Orgaz, Sancho Salcedo, David Camacho. "**A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding**". **Information Sciences**. Vol. 462, pp. 290-314, 2018

Bio-inspired CDAs

- Multi-Objective Genetic Approaches (**Edge-based**)
  1. Generation of communities of edges

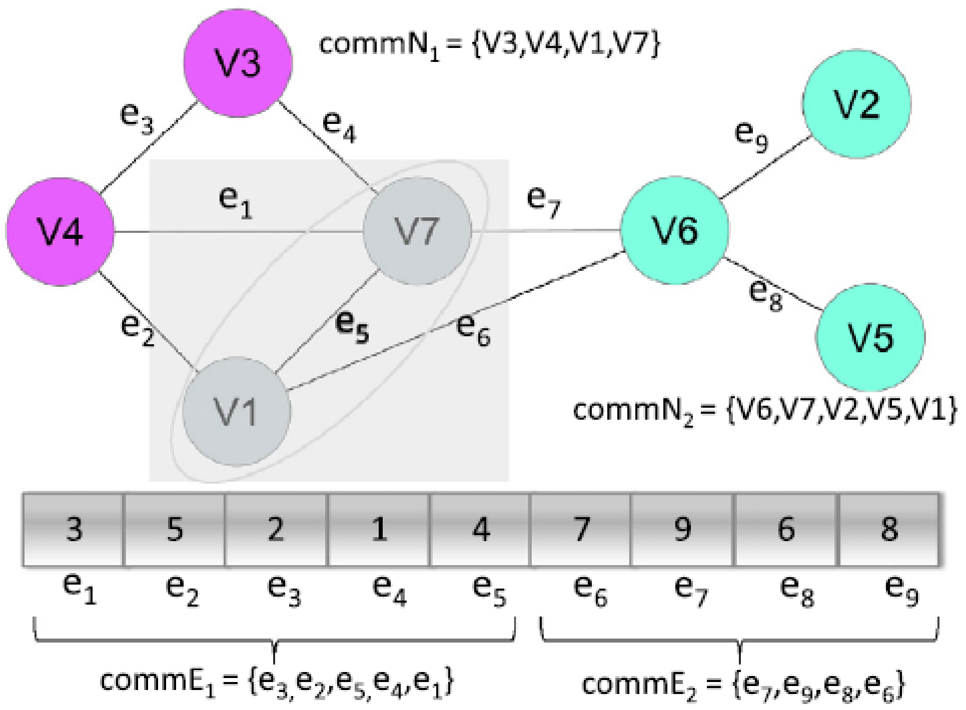




Gema Bello-Orgaz, Sancho Salcedo, David Camacho. "**A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding**". **Information Sciences**. Vol. 462, pp. 290-314, 2018

Bio-inspired CDAs

- Multi-Objective Genetic Approaches (**Edge-based**)
  1. *Generation of communities of nodes*: each edge community is transformed into a node-based community, which contains the **source** and **target node** for each edge belonging to the original community (***decoding phase***)



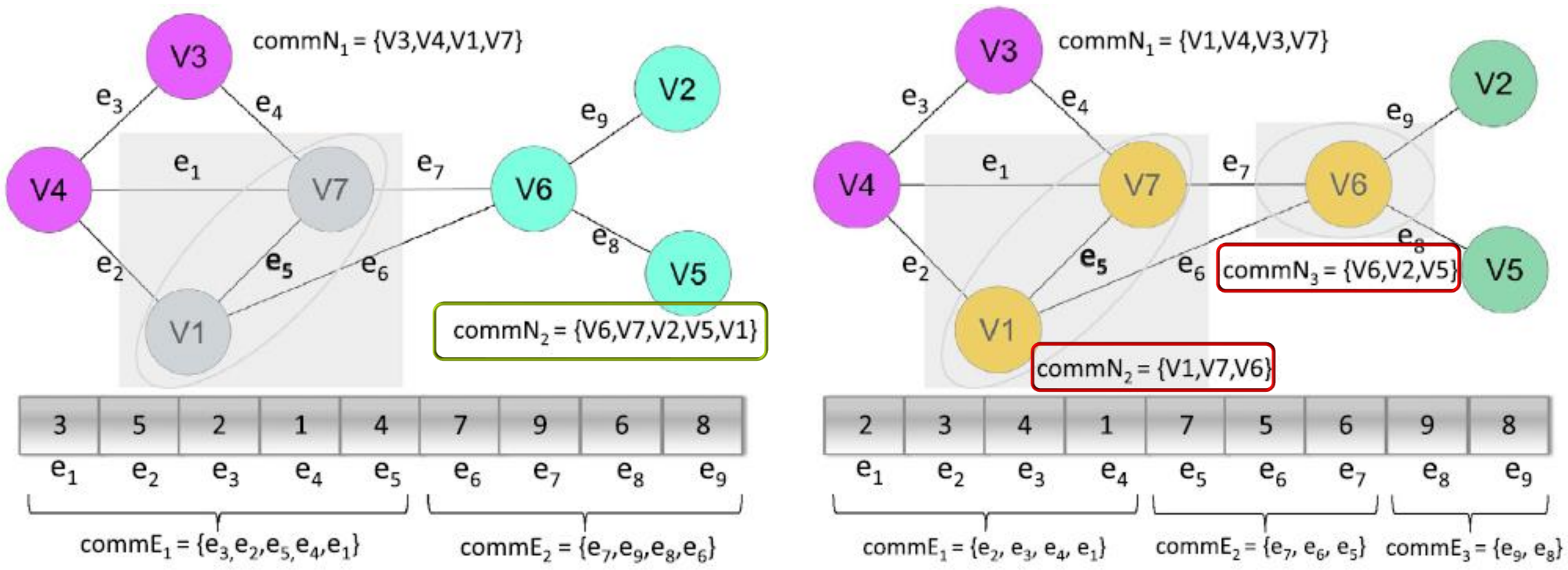
commE1
e3 -> {V3,V4}
e2 -> {V1,V4}
e5 -> {V1,V7}
e4 -> {V3,V7}
e1 -> {V4,V7}
{V3,V4,**V1**,**V7**}: **commN1**

Overlapping nodes:V1, V7

commE2
e7 -> {V6,V7}
e9 -> {V6,V2}
e8 -> {V6,V5}
e6 -> {V6,V1}
{V6,**V7**,V2,V5,**V1**}: **commN2**

Bio-inspired CDAs

- Multi-Objective Genetic Approaches (**Edge-based**)
  1. Generation of communities of nodes



Gema Bello-Orgaz, Sancho Salcedo, David Camacho. “**A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding**”. **Information Sciences**. Vol. 462, pp. 290-314, 2018

Bio-inspired CDAs

- Multi-Objective Genetic Approaches (**Edge-based** vs **Node-based**)

Algorithm Description

The population evolves **according to the node-based version except** for the application of the **three steps**:

1. The **initial population** is randomly generated taking into account that **each allele** has to take a value only between the **adjacent edges**.
2. The **mutation** operator has been modified to **guarantee** that new generated individuals satisfy the **encoding rules**.
3. **Each individual** should be **decode before computing the fitness function**.

Gema Bello-Orgaz, Sancho Salcedo, David Camacho. "**A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding**". **Information Sciences**. Vol. 462, pp. 290-314, 2018

Bio-inspired CDAs: Experimental evaluation

- Two phases:

  1. Analysis of the different metrics related to the network connectivity (internal/external). **Goal**: select the **best metrics to tune up MOCA-OCD** algorithm

  2. MOGA-OCD performance is compared against other traditional CDA algorithms

Gema Bello-Orgaz, Sancho Salcedo, David Camacho. "**A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding**". **Information Sciences**. Vol. 462, pp. 290-314, 2018

Bio-inspired CDAs: Experimental evaluation

- Description of dataset collection of real networks used for the experimental phase

Dataset	Description	Nodes	Edges	GCC	CN	GT	N. Com.
nba_schedule	Games played in the 2013–2014 NBA season	30	421	0.99	5	Yes	6
southernwomen	Southern women social groups	32	93	0	2	No	–
karate	Zacharys karate club	34	78	0.26	15	Yes	2
senate	Senate voting data in 2014	87	1803	0.97	45	Yes	3
football	American football games in 2000	115	613	0.41	9	Yes	12
revolution	Colonial American dissidents	261	319	0	2	No	–
pgp	Interactions in pretty good privacy	10,680	24,340	0.37	25	No	–
p2p–Gnutella25	Snapshots of the Gnutella peer-to-peer file sharing network from August 2002	22,687	54,705	0.01	4	No	–
email-Enron	Enron email communication network (edge indicated that email was exchanged)	36,692	367,662	0.08	20	No	–
brightkite	Friendship network of Brightkite users	58,228	214,078	0.11	37	No	–

Gema Bello-Orgaz, Sancho Salcedo, David Camacho. "**A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding**". **Information Sciences**. Vol. 462, pp. 290-314, 2018

Bio-inspired CDAs: Experimental evaluation

- Comparative assessment of connectivity network metrics using the Karate dataset. **Internal connectivity** metrics are marked in *light gray*, while **External connectivity** metrics are marked in *dark Grey*.

- **Separability** is found as the best metric for the fitness function related to external connectivity

Dataset	Metric m_{ext}	Metric m_{int}	N[C/N]	Den_{avg}	LCC_{avg}	TPR_{avg}	GCC_{avg}	CN_{avg}	Exp_{avg}	Sep_{avg}	CR_{avg}
karate	Sep	LCC	[2.4/20.07]	0.41±0.13	0.73±0.10	0.96±0.03	0.49±0.11	4.48±0.42	1.08±0.52	11.75±5.75	0.06±0.02
		GCC	[2.6/17.44]	0.45±0.17	0.70±0.11	0.95±0.03	0.53±0.17	4.38±0.46	1.48±1.00	9.93±6.08	0.07±0.04
		TPR	[4.4/15.03]	0.49±0.14	**0.75±0.06**	**0.98±0.01**	0.52±0.12	4.23±0.35	1.69±0.83	6.08±3.90	0.08±0.02
		CN	[2/20.65]	0.37±0.16	0.69±0.07	0.95±0.02	0.45±0.12	4.6±0.52	0.89±0.63	**14.86±5.94**	0.06±0.03
		Den	[2.7/15.12]	**0.55±0.10**	0.71±0.11	0.93±0.10	**0.57±0.12**	4.07±0.31	1.70±0.66	8.31±1.94	0.07±0.02
	Exp	LCC	[2/20.5]	0.24±0	0.63±0	0.95±0	0.35±0	**5±0**	0.54±0.03	4.03±0.32	0.04±0
		GCC	[2/20.4]	0.24±0	0.63±0.01	0.95±0	0.35±0.01	4.9±0.21	0.56±0.04	3.91±0.36	0.04±0
		TPR	[2/20.25]	0.24±0	0.63±0.01	0.95±0	0.34±0.01	4.9±0.21	0.57±0.06	3.75±0.47	0.04±0
		CN	[2/20.4]	0.24±0	0.63±0.01	0.95±0	0.35±0.01	4.95±0.16	0.55±0.05	3.93±0.34	0.04±0
		Den	[2.3/21.92]	0.25±0.13	0.62±0.04	0.93±0.06	0.33±0.09	4.66±0.45	0.48±0.27	6.95±3.04	0.04±0.01
	CR	LCC	[2/21.2]	0.23±0.01	0.64±0.01	0.95±0.01	0.35±0.01	4.95±0.16	0.47±0.07	5.52±1.30	0.04±0
		GCC	[2/21.8]	0.23±0.01	0.65±0.01	0.95±0	0.35±0.01	**5±0**	0.43±0.05	6.57±1.08	**0.03±0**
		TPR	[2.2/21.3]	0.23±0.01	0.65±0	0.95±0.01	0.34±0.01	4.93±0.14	0.45±0.06	5.98±1.25	**0.03±0**
		CN	[2/21.9]	0.27±0	0.65±0	0.95±0	0.35±0	**5±0**	0.41±0.02	6.78±0.64	**0.03±0**
		Den	[2/22]	0.23±0	0.65±0	0.95±0	0.34±0	**5±0**	**0.40±0**	7.08±0	**0.03±0**

Gema Bello-Orgaz, Sancho Salcedo, David Camacho. "**A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding**". **Information Sciences**. Vol. 462, pp. 290-314, 2018

Bio-inspired CDAs: Experimental evaluation

- Results of network metrics grouped by **internal connectivity** metrics using ***Separability*** as external function objective in MOGA-OCD algorithm.

- **LCC** is selected as metric for the fitness function related to internal connectivity

DataSet	Metric	Q_{best}	Q_{avg}	NMI_{best}	NMI_{avg}	$FNMI_{best}$	$FNMI_{avg}$
Karate	LCC	0.44	0.35 ± 0.05	0.69	0.52 ± 0.10	0.54	0.42 ± 0.07
	GCC	0.44	0.34 ± 0.06	0.45	0.34 ± 0.09	0.45	0.29 ± 0.13
	TPR	0.42	0.30 ± 0.07	0.44	0.38 ± 0.06	0.45	0.37 ± 0.07
	CN	0.40	0.31 ± 0.11	0.40	0.31 ± 0.08	0.40	0.31 ± 0.08
	Den	0.40	0.31 ± 0.07	0.45	0.34 ± 0.11	0.45	0.31 ± 0.14
Senate	LCC	**0.85**	**0.85 ± 0**	**0.88**	**0.88 ± 0**	**0.70**	**0.70 ± 0**
	GCC	0.85	0.85 ± 0.01	0.88	0.88 ± 0.01	0.70	0.70 ± 0.01
	TPR	0.85	0.85 ± 0.01	0.88	0.88 ± 0.03	0.70	0.69 ± 0.01
	CN	**0.85**	**0.85 ± 0**	**0.88**	0.88 ± 0.01	0.70	0.70 ± 0.01
	Den	**0.85**	**0.85 ± 0**	**0.88**	**0.88 ± 0**	**0.70**	**0.70 ± 0**
Football	LCC	**0.29**	**0.23 ± 0.04**	**0.39**	**0.30 ± 0.06**	**0.32**	**0.26 ± 0.05**
	GCC	**0.29**	0.22 ± 0.05	0.33	0.26 ± 0.06	0.28	0.21 ± 0.05
	TPR	0.15	0.13 ± 0.01	0.36	0.28 ± 0.08	0.25	0.18 ± 0.06
	CN	0.16	0.13 ± 0.02	0.36	0.29 ± 0.07	0.21	0.17 ± 0.04
	Den	0.26	0.32 ± 0.02	0.37	0.27 ± 0.06	0.29	0.23 ± 0.06

Gema Bello-Orgaz, Sancho Salcedo, David Camacho. "**A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding**". **Information Sciences**. Vol. 462, pp. 290-314, 2018

Bio-inspired CDAs: Experimental evaluation

- Comparative assessment of community detection algorithms for all the datasets **with ground truth**.
- Small networks
- CDAs usually provide good results when the graph is **highly structured** and with a small-medium size, so the algorithm can partition it according to its network topology.
- MOGA-OCD obtains similar results

DataSet	Algorithm	N. Com.	Avg. nodes	Q	NMI	FNMI
Nba_schedule	Groundtruth	6	5	0.281	1	1
	CPM	2	15	0.879	0.39	0.24
	Coda	20	5	0.069	0.31	0.19
	Conga	2	15	0.879	0.39	0.24
	Congo	2	15	0.879	0.39	0.24
	MOGA-OCD	2	15	0.879	0.39	0.24
Senate	Groundtruth	3	42	0.810	1	1
	CPM	1	87	0	0	0
	Coda	79	15	0	0.16	0.08
	Conga	2	44.5	0.852	0.88	0.70
	Congo	2	44.5	0.852	0.88	0.70
	MOGA-OCD	2	44.5	0.852	0.88	0.70
Karate	Groundtruth	2	17	0.261	1	
	CPM	3	6	0.515	0.26	0.21
	Coda	36	4	0	0.18	0.10
	Conga	4	8.5	0.476	0.32	0.22
	Congo	3	6	0.322	0.25	0.20
	MOGA-OCD	2.7	17.4	0.358	0.54	0.43
Football	Groundtruth	12	10	0.642	1	1
	CPM	4	13	0.445	0.25	0.16
	Coda	76	7	0	0.45	0.25
	Conga	6	31.5	0.511	0.39	0.24
	Congo	11	9	0.342	0.52	0.49
	MOGA-OCD	4	34	0.216	0.31	0.27

Gema Bello-Orgaz, Sancho Salcedo, David Camacho. "**A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding**". **Information Sciences**. Vol. 462, pp. 290-314, 2018

Bio-inspired CDAs: Experimental evaluation

- Comparative assessment of community detection algorithms for the dataset collection **without ground truth**.
- Large graphs
- For unstructured or sparse graphs, the accuracy and quality of the communities detected by these algorithms significantly decrease (CPM, CONGA,CONGO)
- MOGA-OCD algorithm achieves good results in both quality measures for the different dataset

DataSet	Algorithm	N. Com.	Avg. nodes	Q
Southernwomen	CPM	–	–	–
	Coda	51	4	0
	Conga	4	11	0.475
	Congo	3	14	0.665
	MOGA-OCD	2	26	0.393
Revolution	CPM	–	–	–
	Coda	45	57	0
	Conga	60	5	0.001
	Congo	63	5	0.001
	MOGA-OCD	7	38.7	0.087
pgp	CPM	734	3	0.568
	Coda	100	135	0.560
	Conga	–	–	–
	Congo	–	–	–
	MOGA-OCD	2083	9.3	0.181
p2p-Gnutella25	CPM	540	3.4	0.029
	Coda	98	498.7	0.044
	Conga	–	–	–
	Congo	–	–	–
	MOGA-OCD	2069	26.4	0.040
Email-Enron	CPM	1889	13.9	0.515
	Coda	113	5.04	0.004
	Conga	–	–	–
	Congo	–	–	–
	MOGA-OCD	1448	44.9	0.090
Brightkite	CPM	2098	13.6	0.552
	Coda	109	959.3	0.237
	Conga	–	–	–
	Congo	–	–	–
	MOGA-OCD	2230	47.5	0.101

Gema Bello-Orgaz, Sancho Salcedo, David Camacho. "**A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding**". **Information Sciences**. Vol. 462, pp. 290-314, 2018

New Trends & Challenges

- Related to CDP and bio-inspired algorithms, it can be highlithed three research lines of work:

1. **Temporal networks** (Dynamic behavior): **Dynamic CDAs**, where both nodes or edges can appear or dissapear during the evolution of the network
   - Folino and Pizzuti (2017)
   - Panizo, Bello, Camacho (2018)
2. **Multilayer networks**: where the nodes can be connected by multiple types of relationships
   - Dong et al. (2014)
   - DMultiMOGA (2017)
   - Gonzalez-Pardo, Camacho (2018)
3. **Signed networks**: where the nodes have signed (positive/negative) connections
   - Amelio and Pizzuti (2016)
   - DMultiMOGA (2017)

Publications

1. *"**A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding**". Gema Bello-Orgaz, Sancho Salcedo, David Camacho.* ***Information Sciences.*** Vol. 462, pp. 290-314, 2018
2. *"**Detecting Discussion Communities on Vaccination in Twitter**". Gema Bello-Orgaz, Julio Cesar Hernández-Castro, David Camacho. **Future Generation Computer Systems**, Vol. 66, pp. 125-136, 2017.*
3. *"**ACO-based clustering for Ego Network analysis**". Antonio Gonzalez-Pardo, Jason J. Jung, David Camacho. **Future Generation Computer Systems**, Vol. 66, pp. 160-170, 2017.*
4. *"**Medoid-based clustering using ant colony optimization**". Hector D. Menéndez, Fernando E. B. Otero, and David Camacho. **Swarm Intelligence**, Vol. 10, nº 2, pp. 1-23, May 2016. DOI: 10.1007/s11721-016-0122-5*
5. *"**Combining Social-based Data Mining Techniques to Extract Collective Knowledge from Twitter**". Gema Bello-Orgaz, Hector D. Menedez, Shintaro Okazaki, David Camacho. **Malaysian Journal of Computer Science**. Vol. 27, Issue 2, pp. 95-111, 2014.*

Publications

1. "***Measuring the Radicalisation Risk in Social Networks***". Raul Lara-Cabrera, Antonio González-Pardo, Karim Benouaret Noura Faci, Djamal Benslimane, David Camacho. **IEEE Access**. Vol. 5, pp. 10892-10900, 2017.
2. "***Statistical Analysis of Risk Assessment Factors and Metrics to Evaluate Radicalisation in Twitter***". Raúl Lara-Cabrera, Antonio Gonzalez-Pardo, David Camacho. **Future Generation of Computer Systems** - FGCS. November 2017. DOI: 10.1016/j.future.2017.10.046
3. "***Social networks data analysis with semantics: application to the radicalization problem***". M. Barhamgi, A. Masmoudi, R. Lara-Cabrera, D. Camacho. **Journal of Ambient Intelligence and Humanized Computing**. Online, October 2018. DOI: 10.1007/s12652-018-0968-z
4. "***A new algorithm for communities detection in social networks with node attributes***". H. Gmati, A. Mouakher, A. González-Pardo, David Camacho, D Camacho. **Journal of Ambient Intelligence and Humanized Computing**. Online, October 2018. DOI: 10.1007/s12652-018-1108-5

Bio-inspired Community Finding in Social Networks, new trends and challenges

DAVID CAMACHO

David.Camacho@uam.es

Computer Science Department

Autonomous University of Madrid

Applied **I**ntelligence & **D**ata **A**nalysis
http://aida.ii.uam.es
Universidad Autónoma de Madrid

UNIVERSIDAD AUTONOMA DE MADRID

AI+DA

SUSTech Southern University of Science and Technology