Local Community Computation

Rushed Kanawati
LIPN, CNRS UMR 7030; USPC
http://lipn.fr/~kanawati
rushed.kanawati@lipn.univ-paris13.fr

PLAN

1. INTRODUCTION
2. Community detection
   - Local community detection

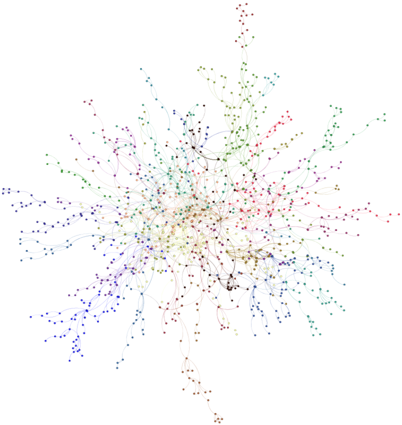
COMPLEX NETWORK

Definition

Graphs modeling (direct/indirect) interactions among actors.

Basic topological features

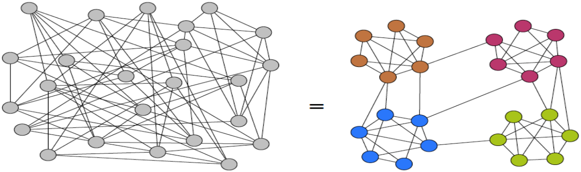
- Low Density
- Small Diameter
- Heterogeneous degree distribution.
- High Clustering coefficient
- Community structure



COMMUNITY ?

Definitions

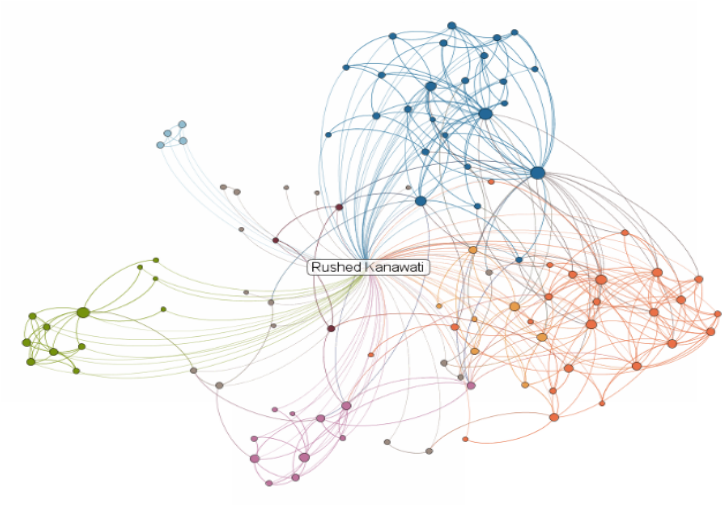
- A dense subgraph loosely coupled to other modules in the network
- A community is a set of nodes seen as one by nodes outside the community
- A subgraph where almost all nodes are linked to other nodes in the community.
- . . .



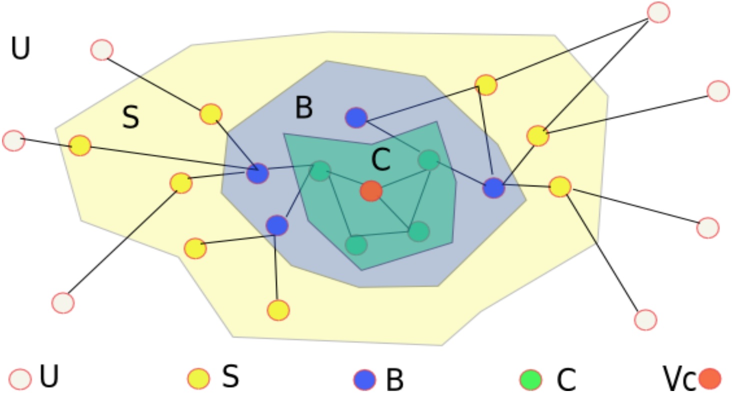
COMMUNITY DETECTION PROBLEM

- Local community identification (ego-centred).
- Network partition computing
- *Overlapping community detection*

LOCAL COMMUNITY



LOCAL COMMUNITY



LOCAL COMMUNITY

1. $C \leftarrow \{\phi\}, B \leftarrow \{n_0\}\ S \leftarrow \Gamma(n_0)$
2. $Q \leftarrow 0$ /* a community **quality function** */
3. While Q can be enhanced Do
    1. $n \leftarrow argmax_{n \in S} Q$
    2. $S \leftarrow S - \{n\}$
    3. $D \leftarrow D + \{n\}$
    4. update B, S, C
4. Return D

QUALITY FUNCTIONS : EXEMPLES I

Local modularity R [Cla05]

$R = \frac{B_{in}}{B_{in}+B_{out}}$

Local modularity M [LWP08]

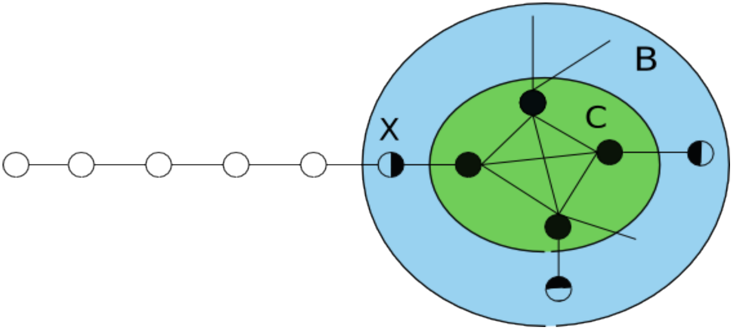
$M = \frac{D_{in}}{D_{out}}$

Local modularity L [CZG09]

$L = \frac{L_{in}}{L_{ex}}$ where : $L_{in} = \frac{\sum_{i \in D} \|\Gamma(i) \cap D\|}{\|D\|}$, $L_{ex} = \frac{\sum_{i \in B} \|\Gamma(i) \cap S\|}{\|B\|}$

And many many others . . . [YL12]

LOCAL MODULARITY LIMITATIONS: AN EXEMPLE



- Blank nodes enhance B_{in} and D_{in} without affecting B_{out} and D_{out}
- Bank nodes will be added if M or R modularities are used
- Low precision computed communities
- **Proposed solution: Ensemble approaches**

Multi-objective local community identification

Three main approaches

- Combine then Rank
- **Ensemble ranking**
- Ensemble clustering

COMBINE THEN RANK

Principle

Let $Q_i(s)$ be the local modularity value induced by node $s \in S$

$$\widetilde{Q_i(s)} = \begin{cases} \frac{Q_i(s) - \min\limits_{u \in S} Q_i(u)}{\max\limits_{u \in S} Q_i(u) - \min\limits_{u \in S} Q_i(u)} & \text{if } \max\limits_{u \in S} Q_i(u) \neq \min\limits_{u \in S} Q_i(u) \\ 1 & \text{otherwise} \end{cases}$$

$$Q_{com}(s) = \frac{1}{k} \sum_{i=1}^{k} \widetilde{Q_i(s)}$$

ENSEMBLE RANKING APPROACHES

Principle

- Rank S in function of each local modularity Q_i
- Select the winner after applying **ensemble ranking** approach
- What stopping criteria to apply ?

Stopping criteria

- *Strict policy* : All modularities should be enhanced
- *Majority policy* : Majority of local modularities are enhanced
- *Least gain policy* : At least one local modularity is enhanced.

ENSEMBLE RANKING

Problem

- Let S be a set of elements to rank by n rankers
- Let σ_i be the rank provided by ranker i
- **Goal: Compute a consensus rank of** S.

ENSEMBLE RANKING

Problem

- Let S be a set of elements to rank by n rankers
- Let σ_i be the rank provided by ranker i
- **Goal: Compute a consensus rank of** S.

Déjà Vu: Social choice algorithms, but . . .

- Small number of voters and big number of candidates
- Algorithmic efficiency is required
- Output could be a complete rank

ENSEMBLE RANKING : APPROACHES

Jean-Charles de Borda [1733-1799]

Borda

- Borda's score of $i \in \sigma_k$:
  $B_{\sigma_k}(i) = \{count(j) | \sigma_k(j) < \sigma_k(i) \, ; j \in \sigma_k\}$.
- Rank elements in function of
  $B(i) = \sum_{t=1}^{k} w_t \times B_{\sigma_t}(i)$.

ENSEMBLE RANKING : APPROACHES

Marquis de Condorcet [1743-1794]

Condorcet

- The winner is the candidate who defeats every other candidate in pairwise majority-rule election
- The winner may not exists

ENSEMBLE RANKING : APPROACHES

Marquis de Condorcet [1743-1794]

Condorcet

- The winner is the candidate who defeats every other candidate in pairwise majority-rule election
- The winner may not exists

Condorcet $\neq$ Borda

- Votes : $6 \times A \succ B \succ C, 4 \times B \succ C \succ A$
- Borda winner : B
- Condorcet winner : A

ENSEMBLE RANKING : APPROACHES

Extended Condorcet criterion

If for every $a \in A$ and $b \in B$, majority prefers a to b the all elements in A should be ranked before any element in B.

Kenneth Arrow, 1921-

Arrow's Theorem

No vote rule can have the following desired proprieties :

- Every result must be achievable somehow.
- Monotonicity.
- Independence of irrelevant attributes.
- Non-dictatorship.

ENSEMBLE RANKING : APPROACHES

John Kemeny 1926-1992

Optimal Kemeny rank aggregation

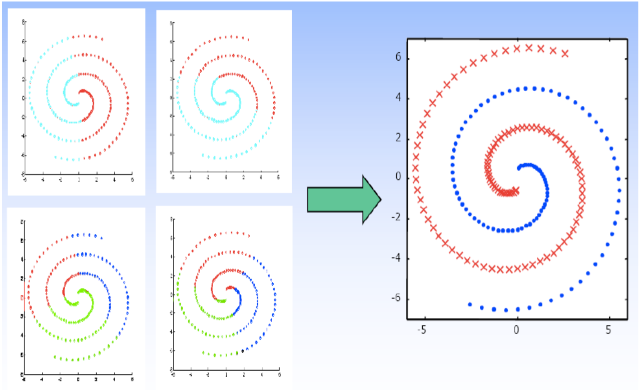
- Let $d()$ be distance over rankings σ_i (ex. Kendall τ, Spearman's footrule)
- **Find** π **that minimise** $\sum_i d(\pi, \sigma_i)$
- NP-Hard problem
- Approximation : Local Kemeny : two adjacent candidats are in the good order.
- **Local Kemeny** : Apply Bubble sort using the *majority preference partial order relationship*
- **Approximate Kemeny** : Apply QuickSort

ENSEMBLE CLUSTERING APPROACHES

Principle

- Let $C_{v_q}^{Q_i}$ be the the local community of v_q applying Q_i.
- We have a natural partition : $P_{Q_i} = \{C_{v_q}^{Q_i}, \overline{C_{v_q}^{Q_i}}\}$
- **Apply an ensemble clustering approach.**

ENSEMBLE CLUSTERING: PRINCIPLE



from A. Topchy et. al. Clustering Ensembles: Models of Consensus and Weak Partitions. PAMI, 2005

ENSEMBLE CLUSTERING: APPROACHES

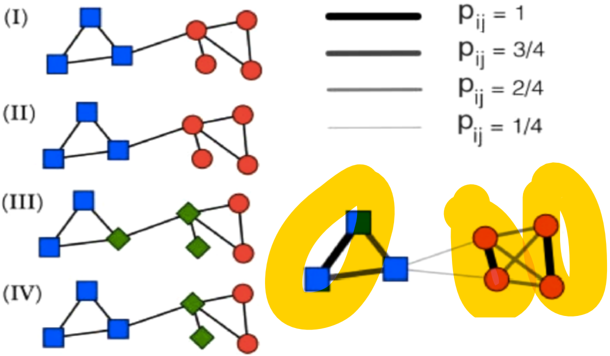
CSPA: Cluster-based Similarity Partitioning Algorithm

- Let K be the number of basic models, $C_i(x)$ be the cluster in model i to which x belongs.
- Define a similarity graph on objects : $sim(v,u) = \frac{\sum_{i=1}^{K} \delta(C_i(v), C_i(u))}{K}$
- Cluster the obtained graph :

  Isolate connected components after pruning edges

  Apply community detection approach
- Complexity : $\mathcal{O}(n^2kr)$: n # objects, k # of clusters, r# of clustering solutions

CSPA : EXEMPLE



from Seifi, M. Cœurs stables de communautés dans les graphes de terrain. Thèse de l'université Paris 6, 2012

ENSEMBLE CLUSTERING: APPROACHES

HGPA: HyperGraph-Partitioning Algorithm

- Construct a hypergraph where nodes are objects and hyperedges are clusters.
- Partition the hypergraph by minimizing the number of cut hyperedges
- Each component forms a meta cluster
- Complexity : $\mathcal{O}(nkr)$

ENSEMBLE CLUSTERING: APPROACHES

MCLA: Meta-Clustering Algorithm

- Each cluster from a base model is an item
- Similarity is defined as the percentage of shared common objects
- Conduct meta-clustering on these clusters
- Assign an object to its most associated meta-cluster

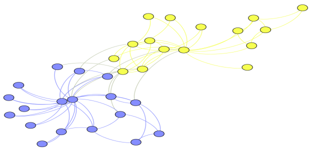
- Complexity : $\mathcal{O}(nk^2r^2)$

Experiments

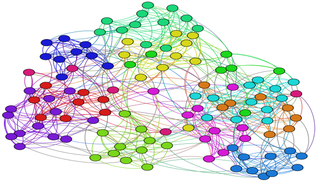
Protocol ([Bag08])

1. Apply the different algorithms on nodes in networks for which a ground-truth community partition is known.
2. For each node compute the distance between the real-partition and the computed one (Ex. NMI [Mei03])
3. Compute average and standard deviation for the network.

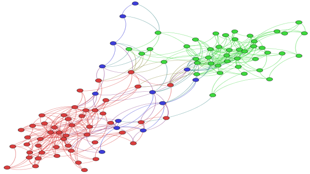
DATASETS



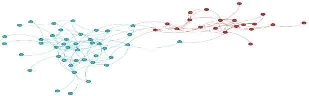
Zachary's Karate Club



Football network

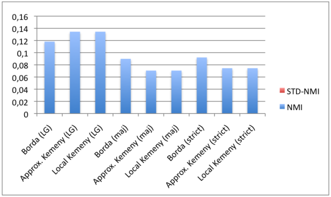


US Political books network

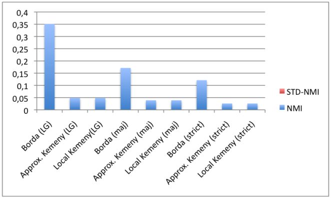


Dolphins social network

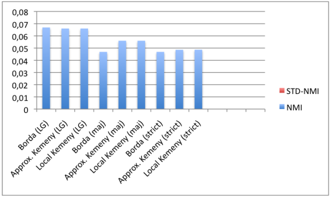
RESULTS : EVALUATING STOPPING CRITERIA (NMI)



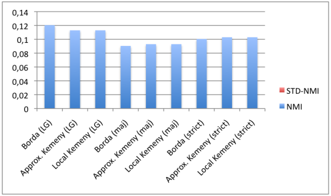
Zachary's Karate Club



Football network

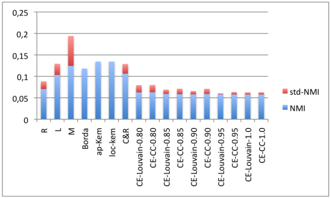


US Political books network

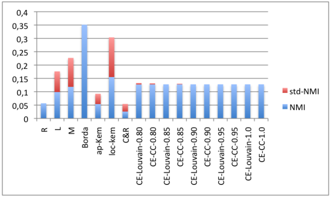


Dolphins social network

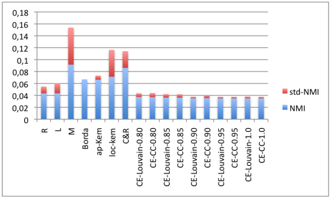
RESULTS : COMPARATIVE RESULTS (NMI)



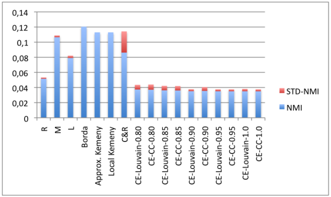
Zachary's Karate Club



Football network



US Political books network



Dolphins social network

BIBLIOGRAPHY I

J. P. Bagrow, *Evaluating local community methods in networks*, J. Stat. Mech. **2008** (2008), no. 5, P05001.

Aaron Clauset, *Finding local community structure in networks*, Physical Review E (2005).

Jiyang Chen, Osmar R. Zaïane, and Randy Goebel, *Local community identification in social networks*, ASONAM, 2009, pp. 237–242.

Feng Luo, James Zijun Wang, and Eric Promislow, *Exploring local community structures in large networks*, Web Intelligence and Agent Systems **6** (2008), no. 4, 387–400.

Marina Meila, *Comparing clusterings by the variation of information*, COLT (Bernhard Schölkopf and Manfred K. Warmuth, eds.), Lecture Notes in Computer Science, vol. 2777, Springer, 2003, pp. 173–187.

BIBLIOGRAPHY II

Jaewon Yang and Jure Leskovec, *Defining and evaluating network communities based on ground-truth*, ICDM (Mohammed Javeed Zaki, Arno Siebes, Jeffrey Xu Yu, Bart Goethals, Geoffrey I. Webb, and Xindong Wu, eds.), IEEE Computer Society, 2012, pp. 745–754.