DEMON
A Local-first Discovery Method For Overlapping Communities

Michele Coscia1, Giulio Rossetti2,3, Fosca Giannotti2, Dino Pedreschi2,3

1 Harvard Kennedy School, Cambridge, MA, US michele_coscia@hks.harvard.edu
2 ISTI - CNR KDDLab, Pisa, Italy {fosca.giannotti, giulio.rossetti}@isti.cnr.it
3 Computer Science Dep., University of Pisa, Italy pedre@di.unipi.it

KDD 2012, Beijing, August 14th 2012
Outline

• Communities and complex networks
  • A matter of perspective

• DEMON Algorithm
  • Properties
  • Experiments
  • Evaluation

• Future Works & Conclusions
Communities

• Communities can be seen as the basic bricks of a network

• In simple, small, networks it is easy to identify them by looking at the structure..
...but real world networks are not “simple”

- We can’t identify easily different communities
- Too many nodes and edges
Are they two different phenomena?

No!
A Matter of Perspective

- The only difference is in the scale
- Locally, for each node the structure makes sense
- Globally, we are tangled in the complex overlap

Idea: a bottom-up approach!
Outline

• Communities and complex networks
  • A matter of perspective

• DEMON Algorithm
  • Properties
  • Experiments
  • Evaluation

• Future Works & Conclusions
Reducing the complexity

Real Networks are Complex Objects

Can we make them “simpler”?  

Ego-Networks

(networks builted upon a focal node, the "ego", and the nodes to whom ego is directly connected to plus the ties, if any, among the alters)
DEMON Algorithm

• For each node n:
  1. Extract the Ego Network of n
  2. Remove n from the Ego Network
  3. Perform a Label Propagation
  4. Insert n in each community found
  5. Update the raw community set C

• For each raw community c in C
  1. Merge with “similar” ones in the set (given a threshold)
     (i.e. merge iff at most the ε% of the smaller one is not included in the bigger one)

Two nice properties

- **Incrementality:**
  Given a graph $G$, an initial set of communities $C$ and an incremental update $\Delta G$ consisting of new nodes and new edges added to $G$, where $\Delta G$ contains the entire ego networks of all new nodes and of all the preexisting nodes reached by new links, then

  $$DEMON(\Delta G \cup G, C) = DEMON(\Delta G, DEMON(G, C))$$

- **Compositionality:**
  Consider any partition of a graph $G$ into two subgraphs $G_1$, $G_2$ such that, for any node $v$ of $G$, the entire ego network of $v$ in $G$ is fully contained either in $G_1$ or $G_2$. Then, given an initial set of communities $C$:

  $$DEMON(G_1 \cup G_2, C) = \text{Max}(DEMON(G_1, C), DEMON(G_2, C))$$

Those property makes the algorithm highly parallelizable: it can run independently on different fragments of the overall network with a relatively small combination work.
Experiments

Networks (with metadata):

- Congress
  (nodes US politicians, connected if they co-sponsor the same bills)

- IMDb
  (nodes Actors, connected if they play in the same movies)

- Amazon
  (nodes Products, connected if they were purchased together)

Compared Algorithms:

- Infomap, non-overlapping state-of-the-art

- HLC, overlapping state-of-the-art
Quality Evaluation – Community size

| Network   | Demon $|C|$ | $|\bar{c}|$ | HLC $|C|$ | $|\bar{c}|$ | Infomap $|C|$ | $|\bar{c}|$ | Modularity $|C|$ | $|\bar{c}|$ | Walktrap $|C|$ | $|\bar{c}|$ |
|-----------|--------|---------|--------|---------|--------|---------|--------|--------|---------|--------|---------|
| Congress  | 425    | 63.3671 | 1,476  | 4.5867  |       |         | 6      | 87.6667| 3       | 175.3333| 7       | 71.8571 |
| IMDb     | 14,004 | 12.6824 | 88,119 | 8.3426  |       |         | 5,991  | 27.1574| 4,746   | 11.9157 | 7,877   | 7.1781  |

Table 3: Statistics of the community set returned by the different algorithms.

- $|C|$ number of communities
- $|\bar{c}|$ average community size
Quality Evaluation - Label Prediction

- Multilabel Classificator (BRL, Binary Relevance Learner)
  - Community memberships of a node as known attributes, real world labels (qualitative attributes) target to be predicted;

<table>
<thead>
<tr>
<th>Network</th>
<th>DEMON</th>
<th>HLC</th>
<th>Infomap</th>
<th>Modularity</th>
<th>Walktrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Congress</td>
<td>0.21275</td>
<td>0.14740</td>
<td>0.00535</td>
<td>0.00099</td>
<td>0.00725</td>
</tr>
<tr>
<td>IMDb</td>
<td>0.44252</td>
<td>0.48078</td>
<td>0.38470</td>
<td>0.10692</td>
<td>0.17488</td>
</tr>
</tbody>
</table>

Table 2: The F-Measure scores for Congress and IMDb dataset and each community partition.
Quality Evaluation - Community Cohesion

• How good is our community partition in describing real world knowledge about the clustered entities?
  • “Similar nodes share more qualitative attributes than dissimilar nodes”

\[ CQ(P) = \frac{\sum_{(n_1, n_2) \in P} |QA(n_1) \cap QA(n_2)|}{\sum_{(n_1, n_2) \in E} |QA(n_1) \cup QA(n_2)|} \]

Iff \(CQ(P) > 1\) we are grouping together similar nodes

<table>
<thead>
<tr>
<th>Network</th>
<th>DEMON</th>
<th>HLC</th>
<th>Infomap</th>
<th>Modularity</th>
<th>Walktrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Congress</td>
<td>1.1792</td>
<td>1.1539</td>
<td>1.0229</td>
<td>1.0373</td>
<td>1.0532</td>
</tr>
<tr>
<td>IMDb</td>
<td>5.6158</td>
<td>5.1589</td>
<td>0.1400</td>
<td>1.4652</td>
<td>0.0211</td>
</tr>
</tbody>
</table>

Table 4: The Community Quality scores for Congress and IMDb dataset and each community partition.
Outline

• Communities and complex networks
  • A matter of perspective

• DEMON Algorithm
  • Properties
  • Experiments
  • Evaluation

• Future Works & Conclusions
Future works

• Extension to weighted and directed networks (completed)

• Parallel implementation

• Modify the merging strategy (in progress)
  • Hierarchical merging
  • ...

• Framework structure
  • i.e. different hosted algorithms that can be used in place of LP to extract communities (according to different definitions)
Conclusions

- DEMON approaches the community discovery problem through the analysis of simpler structures (ego-networks)

- The proposed algorithm outperforms state-of-the-art methodologies

- Possible parallel implementation: high scalability
Thanks!

Questions?

DEMON

Democratic Estimate of the Modular Organization of a Network

Code available @ http://kdd.isti.cnr.it/~giulio/demon/