Online Resourse 4

immediate

March 2, 2019

1 Experimental Results

To experiment with NPCD we select a number of unweighted as well as weighted real world networks. We fix $\rho_0 = 0.4$ and vary θ_0 in the range $\{0.05, 0.10, 0.15, ..., 0.50\}$. For each value of θ_0 we run the algorithm 10 times and record the maximum values of Q_{wo}.

1.1 Ground truth analysis on unweighted networks

To learn how NPCD behaves on networks with ground truth communities we select four unweighted networks namely US-FOOTBALL, POLBOOKS, AMAZON and DBLP.(See Table 1.) Since these networks are unweighted we can compare Q_{wo} with Q_{ov} too. For comparison of Q_{wo} and Q_{ov} in these networks, we record the value of Q_{ov} corresponding to that value of θ_0 for which Q_{wo} is highest. First we consider the US-POLBOOKS network. This network, compiled by Krebs[unpublished], represents books on U.S. politics, with edges connecting pairs of books that are frequently purchased by the same customers of the online bookseller `www.amazon.com` around 2005-06. It has 105 nodes and 441 edges. The nodes were annotated as "liberal", 'conservative' or "neutral"[3]. These three categories denote three ground truth communities of the network. Previous algorithms have detected 2 to 4 communities in the network. Over 10 runs of NPCD for each value of θ_0 in the above specified range, the highest value of Q_{wo} at 0.58 corresponds to $\theta_0 = 0.20$. For this θ_0, Q_{ov} is 0.84. At this value there are 3 communities in the network(see Fig. 1). Of the two major communities one comprises mostly "liberals" and the other one "conservatives". The third community consists nodes from all three categories.

The second unweighted network is US-FOOTBALL. It is a network of US college football teams which were divided into 12 "conferences" where intraconference games were more frequent than the interconference games. The conferences are treated as ground truth communities. The network has 115 nodes and 613 edges. It has been studied extensively in the past and communities ranging from 8 to 13 have been reported[2, 4]. At the highest $Q_{\text{wo}} = 0.63$ corresponding to $\theta_0 = 0.15$, NPCD detects 10 communities in this network(see Fig. 2). Most of the detected communities correspond to single conferences, with just a couple of them corresponding to more than one.

Now we focus on the networks DBLP and AMAZON. The network DBLP is a co-authorship network where two authors(nodes) are connected if they have published a paper together. We ran NPCD on this network for different

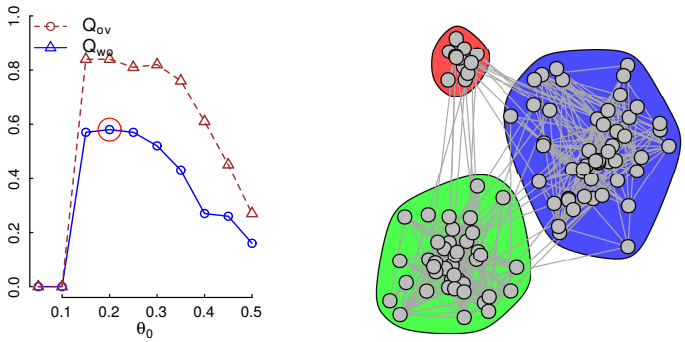


Fig. 1: Community structure in US-POLBOOKS: The left panel plots the maximum values of Q_{ov} and Q_{wo} over 10 iterations of NPCD for each value of θ_0 in the range $\{0.05, 0.10, 0.15, ..., 0.50\}$. The right panel plots the community structure corresponding to the maximum value of Q_{wo} over the entire range of θ_0.

Table 1: Networks studied. Weights and topological characteristics of the networks are as in [1]

Network	Nodes	Edges	Description
POLBOOKS (unweighted)	105	441	This is network of books on U.S. politics, with edges connecting pairs of books that are frequently purchased by the same customers of the online bookseller `www.amazon.com` around 2005-06. The nodes were annotated as "liberal", 'conservative' or "neutral"[3].
FOOTBALL (unweighted)	115	613	It is a network of US college football teams which were divided into 12 "conferences" where intraconference games were more frequent than the interconference games.
NET-SCIENCE-COMP (weighted)	379	914	This represents the giant component of the coauthorship network of scientists working on the network theory and experiment. The unweighted version of the component was previously studied in [2] and recently in [5] for detecting community structure.
HEP-THEORY-COMP (weighted)	5835	13815	This represents the giant component of the collaboration network of scientists who have posted their preprints on the high energy theory archive at `www.arxiv.org` during 1995-1999.
ASTRO-PHYSICS-COMP (weighted)	14845	119652	This is the giant component of the collaboration network of scientists who have posted their preprints on 'astrophysics' archive at `www.arxiv.org`.
COND-MATTER-COMP (unweighted)	36458	171735	This is the giant component of the collaboration network of scientists posting preprints on the condensed matter archive at `www.arxiv.org` during 1995 to 2005.
DBLP (unweighted)	317080	1049866	This is a co-authorship network where two authors are connected if they publish at least one paper together. Publication venue, e.g, journal or conference, defines an individual ground-truth community; authors who published to a certain journal or conference form a community. Each connected component in a group is also regarded as a separate ground-truth community. The ground-truth communities which have less than 3 nodes have been removed.[6]
AMAZON (unweighted)	334863	925872	Network was collected by crawling Amazon website. It is based on Customers Who Bought This Item Also Bought feature of the Amazon website. If a product i is frequently co-purchased with product j, the graph contains an undirected edge from i to j. Each product category provided by Amazon defines each ground-truth community. [6]

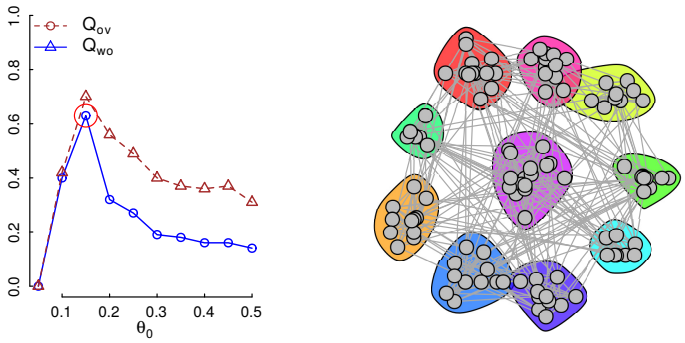


Fig. 2: Community structure in US-FOOTBALL: The left panel plots the maximum values of Q_{ov} and Q_{wo} over 10 iterations of NPCD for each value of θ_0 in the range $\{0.05, 0.10, 0.15, ..., 0.50\}$. The right panel plots the community structure corresponding to the maximum value of Q_{wo} over the entire range of θ_0.

values of θ_0 and found highest value of Q_{wo} as 0.67. The value of Q_{ov} for the corresponding cover is 0.69. Thus both the measures strongly indicate the presence of community structure. It shows that the community structure is significant. The next network AMAZON is a product co-purchasing network where two products(nodes) are connected if they are bought together frequently. For this network when we ran NPCD for different values of θ_0 we found the maximum value of Q_{wo} as 0.71. The value of Q_{ov} for the corresponding cover is 0.48. Again in this case, Q_{wo} strongly indicates the presence of community structure, although Q_{ov} does so only moderately.

1.2 Tests on Weighted networks

Ground truths on weighted networks are not easily available. Therefore, we consider four weighted networks without ground truth as given in Table 1. The first weighted network is NET-SCIENCE-COMP. For this network, in 10 runs of NPCD for each value of θ_0 in the given range, the highest $Q_{\text{wo}} = 0.88$ occurs at $\theta_0 = 0.10$. At this value, the cover has 20 communities with just 0.26% overlapping nodes. The largest community has 78 nodes. The community size distribution, $P(X \leq s)$ – the fraction of communities with size smaller than or equal to s – for this cover follows power law $P(X \leq s) \propto s^{-\alpha}$ with $\alpha = 2.63$(see Fig. 3(a)).

In HEP-THEORY-COMP network, over 10 runs of NPCD, the highest value of Q_{wo} at 0.82 corresponds to $\theta_0 = 0.15$. The cover corresponding to this value of Q_{wo} contains 349 communities with 2337 nodes in the largest community. Only 0.57% nodes have multiple memberships. The community size distribution follows power law with $\alpha = 2.4$(see Fig. 3(b)).

Likewise in ASTRO-PHYSICS-COMP and COND-MATTER-COMP the highest values of Q_{wo} occur at 0.20 and 0.20, respectively(see Fig. 3(c)-(d)). The community size distributions of the covers for these networks also follow power law with $2 < \alpha < 3$ and with less than 1% overlapping nodes.

The results above produced by NPCD are in line with the general characteristics of community size distributions and overlapping and indicate the strong presence of community structure. We also ran NPCD at $\rho_0 = 0.3, 0.4, 0.5$ for the same set of the values of θ_0 and found similar results. The results suggest that keeping ρ_0 in the range $[0.4, 0.5]$ and varying θ_0 in the range $[0.10, 0.30]$ good quality covers can be produced by NPCD. For common users only θ_0 need to be varied keeping ρ_0 at its default value 0.4.

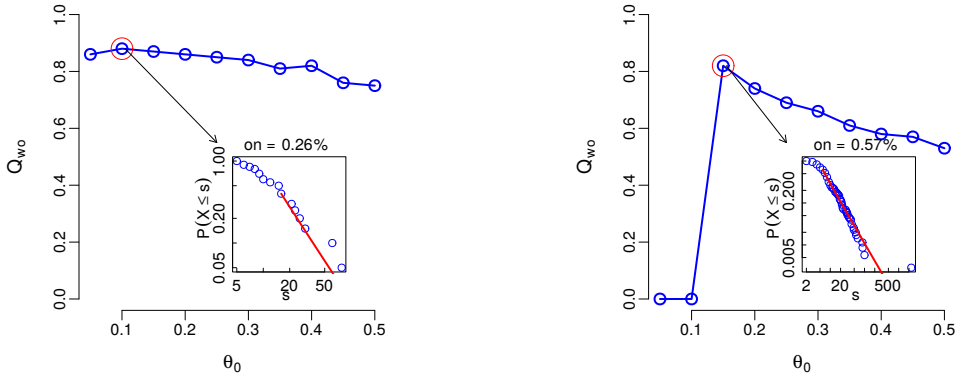
References

[1] M. E. J. Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64(1):016132, 2001. doi: 10.1103/PhysRevE.64.016132.

[2] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74(3):036104, 2006. doi: 10.1103/PhysRevE.74.036104.

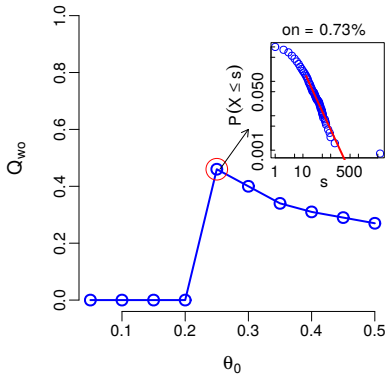
[3] M. E. J. Newman. Modularity and community structure in networks. *PNAS*, 103(23):8577–8582, 2006. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0601602103.

[4] Usha Nandini Raghavan, Rka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76(3):036106, 2007. doi: 10.1103/PhysRevE.76.036106.

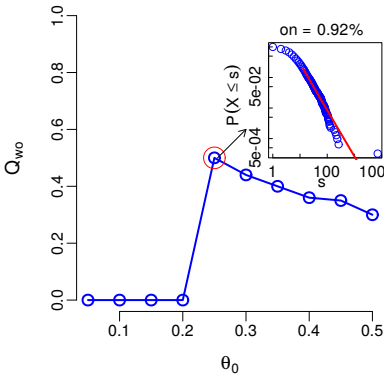


(a) NET-SCIENCE-COMP

(b) HEP-THEORY-COMP



(c) ASTRO-PHYSICS-COMP



(d) COND-MATTER-COMP

Fig. 3: Q_{wo} versus θ_0 graphs for the networks NET-SCIENCE-COMP, HEP-THEORY-COMP, ASTRO-PHYSICS-COMP and COND-MATTER-COMP. The inset figures show community size distributions $P(X \leq s)_{s \geq 1}$, corresponding to the highest Q_{wo} for $\theta_0 \in \{0.05, 0.10, 0.15, ..., 0.50\}$. $P(X \leq s)$ is the fraction of communities that have size smaller than or equal to s. The quantity "on" denotes the number of overlapping nodes corresponding to the highest value of Q_{wo} for the given range of θ_0.

[5] P. Rombach, M. Porter, J. Fowler, and P. Mucha. Core-Periphery Structure in Networks (Revisited). *SIAM Rev.*, 59(3):619–646, 2017. ISSN 0036-1445. doi: 10.1137/17M1130046.

[6] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. *Knowl Inf Syst*, 42(1):181–213, 2013. ISSN 0219-1377, 0219-3116. doi: 10.1007/s10115-013-0693-z.