# Finding community structures in networks by playing pass-the-parcel

**Conor Houghton**

School of Mathematics, Trinity College Dublin, Dublin 2, Ireland

E-mail: houghton@maths.tcd.ie

**Abstract.** Many data sets can be represented by undirected networks. Often, an interesting and important feature of these networks is the existence of communities; groups of nodes whose interconnectivity is higher than the average for the network. Finding these communities can be a difficult problem; exhaustive search and even simulated annealing methods are impractical for larger networks. Here, a different approach is suggested, a measure of the similarity between a pair of nodes is calculated by simulating a game of pass-the-parcel. This similarity is greater for nodes in the same community and so the pass-the-parcel similarity matrix reduces this problem to the better studied problem of clustering. To demonstrate this approach, it is applied to a number of standard data sets. It shows comparable performance to the state-of-the-art extremal optimization and spectral methods. **This algorithm, however, is very similar to one described by Pons and Latapy [1] and so the work described here is not novel.**

PACS numbers: 89.75.Hc,87.23.Ge

## 1. Introduction

An undirected network is a set of nodes where some pairs are connected by links. A number of standard examples are considered in this paper and these serve to illustrate the diversity of network data sets: the standard examples include a sociological study of a university karate club where pairs of members are linked if they are friends and a metabolic network where pairs of substrates are linked if they participate in the same metabolic reaction. Not only are networks of practical importance; the mathematics of network theory has proved to be interesting and rich. One part of network theory that is both useful and mathematically interesting is the study of community structure, the identification of subsets of the nodes with particularly high inter-connectivity.

The *modularity* has been proposed as a measure of how well a network divides into communities [2]. The modularity expresses the idea that the inter-connectivity within a community should be higher than expected. First, the *adjacency matrix* is defined as

$$A_{rs} = \begin{cases} 1 & \text{nodes } r \text{ and } s \text{ linked} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

with $A_{rr} = 0$. The order of a node

$$a_r = \sum_s A_{rs} \tag{2}$$

is the number of links terminating at the node $r$. The total number of links, $m$, is given

$$m = \frac{1}{2} \sum_r a_r \tag{3}$$

where there is a factor of one half because each link joins two nodes. It can be argued that the expected number of links between the $r$ and $s$ nodes is $a_r a_s/2m$ and the modularity $Q$ of a division into communities is

$$Q = \frac{1}{2m} \sum_{r,s} \left( A_{rs} - \frac{a_r a_s}{2m} \right) \Delta_{rs} \tag{4}$$

where

$$\Delta_{rs} = \begin{cases} 1 & \text{nodes } r \text{ and } s \text{ are the same community} \\ 0 & \text{otherwise} \end{cases} . \tag{5}$$

Finding the grouping which maximizes the modularity $Q$ is believed to be NP-hard [3] and so, in practice, some approximate scheme is required. Among the algorithms that have been proposed, two are particularly successful: the extremal optimization method [4] and the spectral method with refinement [5, 3].

The approach taken here is different. Rather than use the network topology directly in searching for community structure, a *similarity matrix* $S_{rs}$ is calculated by simulating a game of pass-the-parcel‡. Consider a node $r$; to calculate the similarity between this node and all the others, a game of pass-the-parcel is simulated in which the node $r$ starts

‡ Pass-the-parcel is a children's party game involving a parcel covered with multiple layers of wrapping, the layers conceal party favours which the children win by removing the layers as the parcel is passed from child to child.

with the parcel. With each iteration, whichever node has the parcel passes it along one of its links, chosen randomly. Roughly speaking, the similarity between the node $r$ and the node $s$ is the number of times $s$ receives the parcel during the game, averaged over repetitions of the game. A parcel will be passed more often within a community than out of it and, so, the similarity between two nodes will be higher if they belong to the same community. This reduces the problem of finding community structure to that of clustering items against a similarity matrix, something for which a number of successful algorithms exist.

## 2. Methods

*Pass-the-parcel*

In the introduction above, the simulated game of pass-the-parcel was described discretely, with the parcel being passed randomly around the network and the similarity calculated by averaging over trials. In fact, it is much more efficient to do a single continuous simulation which calculates the iteration-by-iteration probability density for the parcel: a node passes an equal part of the parcel along all its links. Precisely, each node $r$ has a parcel value, $p_r$, and a cumulative parcel value $c_r$. If, at the $i$ iteration these have values $p_r(i)$ and $c_r(i)$, then the values at the $i+1$ iteration are

$$p_r(i+1) = \sum_{s \in N_r} \frac{p_s(i)}{a_s} \tag{6}$$

and

$$c_r(i+1) = c_r(i) + p_r(i+1) \tag{7}$$

where $N_r$ is the set of nodes linked to the node $r$.

To calculate the whole similarity matrix the game is played once for each node. When it is played for the node $r$, $p_r = 1$ initially and all the other $p_s$'s are zero. The $c_s$'s are also all zero. The game then proceeds though $N$ iterations. If $n$ is the total number of nodes and $a = 2m/n$ the average number of links terminating at each node, $\log_a n$ gives a estimate of the number of iterations required before almost every node has a non-zero $c_r$. In fact, setting

$$N \approx 2 \log_a n \tag{8}$$

appears to be work well here. Now, the $r$ row of the similarity matrix is calculated by normalizing the final $c_s$'s

$$S_{rs} = \frac{c_s}{a_s}. \tag{9}$$

Finally, when the game has been played all $n$ times, once for each node, the similarity matrix is symmetrized in its two indices.

*Identifying communities*

Since there is now a similarity matrix, the nodes can be clustered. The two most commonly used clustering algorithms are probably $K$-means and Hierarchical Agglomerative Clustering (HAC). HAC is used here. Starting with $n$ clusters containing only one node, HAC builds larger clusters by successively joining smaller ones. The similarity of two clusters is calculated as the average similarity between the nodes in the clusters:

$$\text{Similarity}(C_{k_1}, C_{k_2}) = \frac{1}{|C_{k_1}||C_{k_2}|} \sum_{r \in C_{k_1}} \sum_{s \in C_{k_2}} S_{rs} \tag{10}$$

At each step the two most similar clusters are joined. Thus, HAC arranges the nodes into a dendrogram with one cluster at the base, $n$ at the top and, in between, every number in between.

*Refinement*

The problem with HAC is that decisions made early on cannot be changed later and it is common to supplement HAC with a refinement stage. Here, the refinement is chosen to match the Kernighan-Lin inspired refinement introduced in [5]. When HAC is finished, the clustering which has the highest value of the modularity $Q$ is chosen for refinement. For each node $\delta Q$, the change in $Q$ that would result from changing its cluster is calculated. The best change is the one with the highest $\delta Q$; if there is no positive $\delta Q$, this means the least negative one. The best change is performed and the whole process is repeated on the unmoved nodes until every node has been moved once. The intermediate state which has the highest $Q$ value is then selected and the process repeated until it causes no further improvement.

## 3. Results

For ease of comparison, the pass-the-parcel community detection algorithm has been applied to the same networks considered in [4] and [5]. These networks range in size from 34 nodes to 27,519 and the results are tabulated in Table 1. The pass-the-parcel algorithm shows a similar performance to the spectral and the extremal optimization algorithms.

The algorithmic complexity of the pass-the-parcel algorithm is comparable to the spectral and extremal optimization algorithms, for sparse networks this is $O(n^2 \log n)$. Each of the three stages here, the game, HAC and refinement are $O(n^2 \log n)$. However, memory requirement is a significant weakness of the pass-the-parcel algorithm. For large networks, it would be expensive to store the whole $n \times n$ similarity matrix.

To calculate Table 1 the values were thresholded to make the $S$ matrix easier to store. For each row, the diagonal element is set to zero, the row average is calculated and all values less than this are set to zero. After thresholding, the number of nonzero values ranged from 0.044 to 0.323 for the networks considered here. For the first four

**Table 1.** Comparing different algorithms for detecting community structure. Five algorithms are compared here; as well as the extremal optimization (**DA**) and spectral (**N**) algorithm mentioned earlier, it includes **GN**, a pioneering algorithm which finds the links which carry the largest number of shortest paths and eliminates them [2] and **CNM**, a very quick greedy algorithm which is well suited for extremely large networks [6]. These are applied to six data sets of varying size and structure. These six sets have become something of a standard for comparing community identification algorithms. They are the **karate** network [7], a network of **jazz musicians** where the musicians are linked if they played in the same band [8], the **metabolic** network [9, 4], an **email** network based on communication in a medium sized university [10], a PGP web of trust (**key signing**) [11, 12] and a network of **physicists** who have placed papers on the cond-mat arXiv, the links represent co-authorship [13]. Apart from the addition of the results calculated using the pass-the-parcel algorithm, this table is substantially based on one appearing in [5] and is similar to one appearing in [4].

| | | Modularity $Q$ | | | | |
|---|---|---|---|---|---|---|
| network | size $n$ | GN | CNM | DA | N | Here |
| karate | 34 | 0.401 | 0.381 | 0.419 | 0.419 | 0.419 |
| jazz musicians | 198 | 0.405 | 0.439 | 0.445 | 0.442 | 0.444 |
| metabolic | 453 | 0.403 | 0.402 | 0.434 | 0.435 | 0.433 |
| email | 1133 | 0.532 | 0.494 | 0.574 | 0.572 | 0.575 |
| key signing | 10 680 | 0.816 | 0.733 | 0.846 | 0.855 | 0.864 |
| physicists | 27 519 | – | 0.668 | 0.679 | 0.723 | 0.730 |

networks, it has been verified that the thresholding makes very little difference to the result.

## 4. Discussion

When applied to the six standard networks, this new algorithm finds community structures with similar modularities to those found using the extremal optimization and spectral algorithms. It is interesting that these three algorithms are all very different and, indeed, there are a large variety of community detection algorithms [14] including a random walk method [15]. This random walk method differs from our algorithm in that it measures the average length of a random walk between two points.

The pass-the-parcel algorithm has a pleasingly intuitive motivation; it does suffer, however, from being rather ad hoc. In fact, although the other two algorithms are better motivated mathematical, none could really be described as well understood from a mathematical point of view and interest in the mathematics of community detection is likely to be ongoing. Although the need to store the similarity matrix is a disadvantage to the algorithm in its current form, it may prove useful in practical applications because it is well suited to parallel, distributed or on-line implementations.

**Acknowledgments**

Acknowledgements Science Foundation Ireland grant 08/RFP/MTH1280 and support by the Mathematics Applications Consortium for Science and Industry are acknowledged. I am grateful to Alex Arenas and Mark Newman for permission to use the data supplied on their websites.

## References

[1] P. Pons and M. Latapy. Computing communities in large networks using random walks (long version). *Journal of Graph Algorithms and Applications*, 10:191–218, December 2005.

[2] Michelle Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99:7821–7826, 2002.

[3] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74:036104, 2006.

[4] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72:027104, 2005.

[5] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582, 2006.

[6] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[7] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

[8] P. Gleiser and L. Danon. Community structure in jazz. *Advances in Complex Systems*, 6:565, 2003.

[9] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabsi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[10] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Physical Review E*, 68(6):065103, December 2003.

[11] M. Boguña, R. Pastor-Satorras, A. Diaz-Guilera, and A. Arenas. Models of social networks based on social distance attachment. *Physical Review E*, 70:056122, 2004.

[12] X. Guardiola, R. Guimera, A. Arenas, A. Diaz-Guilera, D. Streib, and L. A. N. Amaral. Macro- and micro-structure of trust networks. *arXiv:cond-mat/0206240*, 2002.

[13] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98:404, 2001.

[14] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *J. Stat. Mech.*, page P09008, 2005.

[15] Haijun Zhou. Distance, dissimilarity index, and network community structure. *Phys. Rev. E*, 67(6):061901, Jun 2003.