STAT 517
Ryan Heiderman
11-15-2018

Literature Review

Structure and inference in annotated networks
M. E. J. Newman and Aaron Clauset
Published 14 July 2015 in Nature Communications

"We focus in particular on the problem of community detection in networks and develop a mathematically principled approach that combines a network and its metadata to detect communities more accurately than can be done with ether alone." Newman and Clauset

Introduction

Networks provide a powerful and compact representation of the internal structure of complex systems. Network analysis aims to characterize the structural features and the behavior of the system. (Newman. 2010) Networks are sets of nodes and their interactions. Each node can be accompanied by annotations and metadata. Community detection searches for division of nodes into groups or classes. The divisions can be assortative (grouped by similar characteristics) or disassortative (more connections between groups than within). This paper focuses on the community detection problem, also known as node clustering or classification. The aim is to use metadata to improve that accuracy of community detection without assuming prior relationships of the metadata with those communities. The method also seeks to select between competing divisions of a network by incorporating metadata that correlates with a particular node division. Just as well it seeks to determine if no correlations between network structure and metadata exist.

Examples of Application

A well-known community detection example is the Zachary Karate Club social network. (Zachary, 1977) In this example, a karate club on a university campus consisting of 34 members had a social conflict arise, leading to the group splitting in half. The two at the middle of the conflict were the instructor and the club president, who after the split, each started a new club. Using information regarding communication and friendship between the members, Zachary was able to correctly predict which new group the members would join after the split. This is an early example of using metadata to predict communities within a network. The Zachary Karate Club is a small network, consisting of only 34 nodes. "Community detection algorithms have typically been tested on a low number of real networks where classification of the nodes is available." (Hric, Darst and Fortuanato, 2014). As larger, more complex datasets have become available, testing of the old algorithms have performed poorly.

Some complex systems represented by networks include technological or informational networks such as the internet or online social networks, others include biological such as molecules, cells and food webs. Network analysis may attempt to determine the most influential or central individual within a social network, or find pathways in metabolic networks to better understand the molecular machinery of a cell. When looking at the metadata of a social network it could include such information as the person's age, gender or ethnicity, as well as the data capacity or physical location of the node on the internet. These divisions and correlations between network and metadata may allow the prediction of community membership for nodes which lack network data and only have metadata. Using this approach, community groupings may be predicted by metadata alone without assumptions of the network connections.

Methods

The model generates a network based on each node's probability of belonging to a community depending on the node's metadata. Once every node is assigned a community, edges are randomly and independently placed. The method uses a modified version of a stochastic block model, which uses techniques of Bayesian statistical inference. A standard stochastic block model partitions nodes of a network into subgroups with distribution of ties between nodes dependent only on the block to which the nodes belong. Unfortunately, simple stochastic block models do not perform well when applied to complex, real-world networks. (Karrer and Newman. 2010)

The authors seek to improve and modify the stochastic block model for better community detection in complex, real-world networks. It is modified in two particular ways for network detection, the use of a degree-correction factor which matches the number of predicted connections each node has with observed data and the introduction of a dependence on node metadata, which sets a probability of a node belonging to a certain community. The model makes use of an expectation maximization algorithm to place a connection. The algorithm finds the optimal division of the network into communities based on a given network's adjacency matrix, prior probability, marginal posterior probability and joint probability. The full distribution of probabilities becomes exponentially large given more and more nodes, this paper proposes a method of calculating it based on belief propagation.

The metadata serve to determine the prior probability of a node belonging to a particular community. For an undirected network of n nodes divided among k communities, each node could have discrete, unordered values as metadata, or ordered, and potentially continuous variables as metadata. Metadata may also not necessarily be one dimensional. For example, metadata of a social network noted could be two dimensional describing both language and race (e.g. Spanish/white or English/black). Metadata could also be missing and the metadata value simply becomes 'missing'.

A network is generated as follows, each node is assigned a community based on the probability depending on the nodes metadata. Then after every node has been assigned to a community, edges are placed independently and at random between nodes with the probability of an edge based on specified parameters and arbitrary degree sequences. Then community detection is applied to this network by fitting based on maximum likelihood. There is a marginal posterior probability that a node belongs to a particular community, and a joint probability the two nodes bellowing to two communities respectively. The prior probabilities can then be used to determine how and to what extent the metadata are correlated with the communities.

The expectation maximum algorithm always converges to a maximum of likelihood, but is not guaranteed to be the global maximum. So it will be necessary to repeatedly apply the algorithm with different random initial guesses, then the run with the highest final value of likelihood is determined and used for the final fitted values. In this paper they used at least ten random starts for each network.

To quantify the extent to which the communities found by the algorithm match the 'real-world', or ground truth, the normalized mutual information (NMI) criteria is calculated. NMI range from 0 to 1, with 0 meaning the metadata are uninformative and 1 meaning the metadata predict communities perfectly. NMI makes use of conditional entropy, which is equal to the amount of additional information needed, on top of the metadata alone, in order to specify the community membership of every node in the network. So there is a probability that a node contains a certain metadata, then there is the probability that a node will belong to a particular community given that it contains that particular metadata. If the metadata perfectly describe community division, the conditional entropy is 0.

Results

The model was applied to a range of synthetic and real-world networks to test the models ability to detect network structure.

Synthetic Network

The first test was on computer-generated (synthetic) networks which know community structure using a technique known as a Planted Partition Model. The networks were created with standard stochastic block model where nodes are assigned to groups with edges placed between them independently with probabilities that are a function of group membership only. Then after the synthetic network was created, discrete-valued node metadata was generated at random to match the true community assignments of nodes, as well as non-matching metadata values. The generation of matching/non-matching metadata was varied to control the extent to which the metadata correlated with community structure to allow tests on the algorithms ability at different levels. Levels of metadata matching ranged from 50% to 90%. Community structure was also varied by controlling the within-group and between-group edge probabilities. A strong community had more within group edges than between, and a weaker community was one where the amount of within group edges approached the amount of between group edges.

As expected, when correlation between metadata and community agreed the algorithm's performance was strongest. When the metadata and community agreed for exactly half of the nodes, there was no correlation and thus the model using metadata was not helpful in community detection. Another clear pattern was that for strong community structure (many within group edges) the algorithm performed strongly and classified essentially all nodes into correct groups. When community structure was weakened the algorithm performance dropped, but even with a weak community structure a high correlation with metadata helped the algorithm's performance. "The fact that our algorithm does better when there are metadata thus implies that the algorithm with metadata does better than any possible algorithm without metadata." This statement is supported by the fact that the algorithm still correctly classified a fraction (relative to the fraction of match between metadata and communities) of the nodes below the 'detectability threshold' (where community structure becomes so weak it is undetectable by any algorithm that relies on network structure alone). The detectability threshold is shifted, the authors suggest, and maybe could be eliminated completely with strongly correlated metadata.

A different synthetic test was performed on the algorithm's ability to detect competing divisions of a network. The test was to divide four equally sized communities into two groups, resulting in eight possible choices. The test was run 100 times between the metadata algorithm and a classic algorithm containing no metadata. Given only a 65% correlation between metadata and community the algorithm found the correct division of the network 98% of the time, by contrast without metadata the correct division was found only 6% of the time.

Real-world Networks: Student Friendships

The next tests of the algorithm were performed on real-world networks of varying community types. The first application was on friendship networks at an American high school and its feeder middle school. The sample size was 795 students of ages 12 to 14 and 14 to 18, in grades 7 through 8 and 9 through 12 respectively. The network represents patterns of friendship established by survey. The algorithm was asked to divide the network into two communities, selecting division by using metadata that correlate with the one of interest. The first test used the six school grades as metadata and the algorithm correctly identified the division between middle school and high school. The next test used the students' self-identified ethnicity as metadata. The algorithm found a completely different division into two communities than the grade information. This division was a group of white students and another of black students with remaining smaller groups of ethnicities distributed roughly evenly between the groups. The last test of the algorithm was using gender as metadata. This time the division did not divide into the possibly expected division into a group of male and a group of female but instead found a hybrid of grade and ethnicity. Which means the algorithm basically ignored the gender metadata because there wasn't a good network division to be made that correlated with it. This is because the algorithm only makes use of the metadata when it improves the network division quality. For the school friendship network algorithm tests run with grade and ethnicity metadata, the NMI scores were 0.881 and 0.820 respectively, suggesting strong prediction but for the gender metadata the NMI score was 0.003 indicating the gender metadata had essentially no influence on community prediction.

Real-world Networks: Predator-prey Interactions

The next network the metadata prediction algorithm was applied to was an ecological one. A food web of predator-prey relationships amongst 488 marine species in the Weddell Sea, part of the Southern Ocean out the coast of Antarctica. There were many metadata attributes available for the species including feeding mode (which is whether a species was a deposit feeder, suspension feeder or a scavenger), the zone of the ocean the species lives and others, including body mass. Body mass was the focus of the real-world test on this ecosystem. This metadata is considered an ordered, continuous variable. Since there was such a wide range in body mass, from microorganisms up to wales, the mass data was logarithmically transformed, and k-way community decompositions using this log-mass data was performed. K-way community detection is a hierarchical analysis of the community structure. The algorithm was applied using k=3 and the divisions found match the ecological roles closely, with one group composed almost entirely of primary producers and herbivores, the next group with primarily omnivores and the last group containing most of the carnivores. The node sizes, indicating larger body mass, increase as you move from group 1 to 3, hitting on the well-known correlation between body mass and ecosystem role. An application of this community detection problem is shown with the introduction of a new species with unknown community membership. The probability of belonging to a group is highly correlated with body mass, and a new species with very low body mass has a high probability of belonging to group 1, meaning it is most likely a primary producer or a herbivore.

An interesting result of this application was the prior probabilities found as a function of body mass. They recovered a well-known correlation between body mass, trophic level and ecosystem role. That is, larger animals tend to be carnivores, and smaller animals tend to be herbivores or primary producers. These roles were not part of the metadata, only body mass, yet the result found by the algorithm was that of the ecological roles of these animals.

Real-world Networks: Internet Graphs

The third real-world application of utilizing metadata for community detection involves peering structure nodes of the internet. The dataset included over 46,000 node representations of the internet at the level of autonomous systems, with metadata being location of system as the country it is housed. First a traditional expectation maximization algorithm was used to 'blindly' divide into 5 communities, followed by application of the metadata algorithm. NMI was used again to determine if metadata increased the predictive probability. In detecting network communities via EM algorithm, NMI values ranged from 0.398 to 0.626. With the addition of country as metadata to determine network division, NMI value were significantly higher at 0.870.

Real-world Networks: Facebook

The algorithm was applied to an early facebook friendship network of more than 15,000 students at Harvard University. The nodes represent students and the edges represent friend relations on facebook. Metadata of several types was incorporated at each node including gender, graduation year, major and the dorm the person lives in. The metadata provided the learned prior probability of community membership and found strong correlation between communities and graduation year, and dormitory in particular. The authors note this is unsurprising considering how friendships are formed at college, most likely your friends are those in your class and around you in your dormitory. The NMI values for graduation year and dorm are 0.668 and 0.255 respectively. These NMI values increased from the blind network detection, implicating the value of including metadata in this network analysis.

Conclusion

The authors described a technique for directly incorporating metadata into the analysis of networks, in particular the problem of community detection. They showed that the incorporation of metadata can improve the accuracy of community detection amongst a range of network types including social, biological and technological. The algorithm is flexible in that it will automatically choose whether to use or ignore the metadata in deciding divisions. The authors suggest further work including more complex metadata types and other community detection problems such as hierarchy and ranking, and also to predict missing metadata in incomplete datasets.

References

- M.E.J. Newman and A. Clauset. 2015. Structure and inference in annotated networks. Nature Communications.
- M.E.J. Newman. 2010. Networks: An Introduction. Oxford Press.
- D. Hric, R.K. Darst and S. Fortunato. 2014. Community detection in networks: Structural communities versus ground truth. Phys. Rev.
- B. Karrer and M.E.J. Newman. 2010. Stochastic blockmodels and community structure of networks. Phys. Rev
- W.W. Zachary. 1977. An information flow model for conflict and fission in small groups. Journal of Anthro Research