Rabiya Hozar
504161548

Modelling a disease-relevant contact network of people who inject drugs

David A. Rolls, Peng Wang ,Rebecca Jenkinson ,Phillipa E. Pattison ,Garry L. Robins ,Rachel Sacks-Davis ,Galina Daraganova ,Margaret Hellard, Emma McBryde

Introduction

This paper aims to model a contact network of people who inject drugs and show the relation between drug injection and Hepatitis C disease spreading. By this way result of this paper can help to understand Hepatitis C spread and infect tramission flow in further researches and develop prevention method. To form exponential random graph (ERGM) location, age, injecting frequency and gender is used as parameters. As a result, a novel model-dependent estimate of network size is generated. To evaluate result of research, comparison is made between Bernoulli graphs and this novel approach. Previous papers show that node degree distribution, in other words distribution of number of contact per person, is key point to analyze epidemiology of Hepatitis C. But these papers do not mention importance of clustering. For this novel approach clustering also is taken cared.

Methods

Data Collection

Data to form network is collected using snowball sampling in Melbourne, Australia. A snowball sample is a non-probability sampling technique when people in target population is hardly to find directly. Researchers identified few member of population then ask them if they know other members of same population. In this data collection basically people are interviewed by researchers and asked to name 5 other people who inject drugs. They are asked about injecting frequency, tendency to share needles or syringe. Plus, a HCV test is done to identify if person has disease.

Network Construction

Collected data contains 258 people with at least one identified partner and 47 other people that researches was not able to interviewed with their partners. Two network type is suggested. One of them is “S” type where there is a connection between X and Y; Y uses any equipment such as needless or syringe that is used by X before. ”S” states for “sharing”. Other network type is “U”, In network “U” there is a connection between X and Y if they inject drugs in same place and time. “U” states for “use with”. In sampled data S is very infrequent so network type, thus “U” is chosen and it is undirected graph. There are missing network data in sample because some of the people who injecting drugs cannot be interview since they reject it or cannot be found. For proper result network component size is determined as three and above. Since number of nodes that are isolated, because their partners cannot be found, is high. Main focus of research is simulating the network to identify Hepatitis C transmission through injecting drugs.

ERGM

Conditional estimation and seed set reconstruction

Collected data is actually just a sample, it is not a complete network. EGRM cannot be applied on sample data, because there must be a complete network to apply EGRM. Thus SnowPNet is implemented in this research. SnowPNet uses Pattison et al. (2012) technique. So that Zone structure can be achieved. According to Zone structure, these initial nodes are in Zone 0. Nodes that are connected to nodes in Zone 0 is in Zone 1. Basically nodes that are connect to Zone k are in Zone k+1 Important rule is there should not be any connection between not adjacent zones such there is no connection between Zone h and Zone m if $|h-m| \geq 2$.

As stated clustering is important for this research's approach. So in addition to 4 main attributes (location, age, gender, injecting frequency) edge, isolation, k-stars, k-triangle, alternating k-2-path are used. Also four main attribute for simplicity all defined as boolean. If age is smaller than 25 this means 1(yes) otherwise 0(no). Location can be 1,2,3. For injecting frequency daily usage takes 1(yes)

other frequencies are 0(no). Lastly for gender 1 represent female, 0 represent male. After estimation, results show that location is the largest positive estimate. Other significant parameters are edge, isolates, k-triangles, k-2-paths, same age, daily user frequency. So paper's EGRM show that gender does not have importance.

Attribute modelling

To be able to simulate network size for a node all possible attributes must be provided. (3 locations, 2 genders, 2 frequency option and age, it makes 24 possible combinations in total). Study show that 3 parameters combinations also fit data, thus combination is used. Joint distribution location and age, distribution of gender and distribution of daily user frequency. Data is collected as directed but modelled graph is undirected. Because there can be ambiguity in zone membership. There can be doubt in zone membership if two people identify each other but interview time between two participants is not enough for time resolution of recorded data, because of multi-wave sampling. In other words, researches made interview too soon. Or participant identified each other causes cycle.

Results

Estimation of model parameters are observed using SnowPNet, Markov Chain Monte Carlo Maximum Likelihood Estimation. Nodes in zone 3 and beyond are excluded. In previous papers obtaining Fig 1 based on zone structure. But there are too many number of nodes in zone 0 and nodes are too close, because of snowball data can be biased. To avoid this, paper suggests that seed nodes must have distance at least three from each other. So this satisfies that each edge is in the neighborhood of single seed node and result with non-overlapping neighborhood. Each of the edge is associated with only single seed node. Using this rule 20 seed set are formed. And other nodes were assigned to other zones.

Community-size determination

From original data with 49 seed nodes and zone 0-2 is network models are simulated to obtain network size, node choose is made randomly. 100 repetitions are made and selected confidence interval rate is 95%. Each obtain network is compared using 9 parameters that stated. Result can be seen in Fig.3 in original paper. For nine parameters estimated network size fits to 524.

Community-size simulations

48000 network is simulated with size 524 using PNet. Largest component of each network is taken because aim of this research is to evaluate treatment intervention strategies to reduce hepatitis C spreading among drug injection people. So most connected and the one with longest path is chosen. Small component is considered as isolated. Fig 6 shows the one of largest component. As stated it has long paths. Result of EGRM is compared with Bernoulli graphs using largest component of two models. Clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. Test show that EGRM results better. EGRM has better clustering coefficient, highly clustered. Based on geodesic length (shortest path), results are showing that EGREM has longer paths. Another comparison point is cut points. Cut points are the nodes that if you delete node it creates fractures in network. If some has preventions such as vaccine creates cut point in network. Since node does not belong any path of infection. Removing cut points creates minimum component size. EGRM has higher probability of randomly chosen point is cut point. This means if any prevention methods such as vaccine is applied for a random node, it reduces the probability of infection. Also EGRM shows that people in location 3 and younger 25 have more probability to be cut point.

Discussion

As conclusion, main objective of this paper is to the develop a social circuit of PWID (people who injected drugs) that show relevance to HCV infection. As a novel approach it uses conditional estimation. Mainly there are 4 attributes from sample data to model graph using homophily age,

location and daily user determined as significant attributed. After estimation EGRM show that gender was less significant. Papers also includes clustering parameters such as k-triangles and k-stars to achieve highly connected network. Graph is undirected graph. So it shows use with as stead before. Using undirected graph make it less sensitive to flawed recall. Directed graphs can be used to show who used after whom. Since data is collected using snowball sampling, there are some points to consider. For example, in snowball based methods initial seed are highly effective on all data set. Thus in this paper while selecting seed nodes a distance rule is applied. As a result, estimated network size 525 to nine parameters. Interview people identify at most 5 people in data collection. But there is a chance someone can be identified more than 5 times. Thus node degrees can be larger than 5. It effects the node component size. For further research, a proper study should be made for bias, consistency for estimating network size technique. Also there were missing nodes in sampling data. They are taken as unused seed nodes. This approach can be improved.