Rumor source inference

Temigo Aoiki

April 30, 2015

This is a reference document for whisper project.

Contents

1 Overview **2**

2 Epidemic models **2**

2.1 The SIR model . . . 2

2.2 The SI model . . . 2

3 Rumor centrality **2**

3.1 General assumptions . . . 2

3.2 Regular trees . . . 3

3.3 Computing rumor centrality . . . 3

3.4 General trees . . . 4

3.5 General graphs . . . 4

4 Multiple sources **4**

5 Suspects **4**

6 Limited observations **5**

7 Related projects **5**

1 Overview

Research in epidemiology has been focused on the dynamics of epidemics, infection rates, network structure, etc. However, only very recent work studied the source inference of an epidemics.

In this document, we will will work on graphs that are supposed to be given. (Graph -¿ Source inference) The modelisation part (Real world -¿ Graph) won't be studied here.

2 Epidemic models

The spread of information should first be designed. [1] Here, we only present the SIR model and the SI model, which are the main models studied in the rumor source inference literature.

2.1 The SIR model

The well-known model SIR model (susceptible-infected-recovered) is the most popular. It was invented in 1927 by Kermack and McKendrick. Population is partitioned into three compartiments :

- S the susceptible individuals (not yet infected),
- I the infected population,
- R the individuals who are no more concerned by the disease (either because they recovered and are immune, or because they died).

The sum $N = S+I+R$ remains constant (it is the total number of individuals). Considering e.g. rates of infection and recovery, we can then define several differential equations on these functions.

2.2 The SI model

A variant of SIR is the SI model (susceptible-infected), where there is no possible recovery.

3 Rumor centrality

Shah and Zaman provided the first systematic study of the problem of finding the source of a rumor in a network. They defined the topological quantity of **rumor centrality** .

3.1 General assumptions

Let $G(V, E)$ be an undirected graph modelling the network. V is the set of vertices, E is the set of edges.We assume there is only one rumor source v^*. The model used here is the SI model.

[1] Good deal of informations on the different models used in epidemiology can be found here : http://en.wikipedia.org/wiki/Epidemic_model

We define the virus graph G_N which is the subgraph of G induced by the N infected nodes. The actual source will be denoted v^* and the source estimation will be $\hat{v}$.

Each node is equally likely to be the source. Thus, the Maximum Likelihood Estimator is the best estimator, that is :

$$\hat{v} = \arg\max_{v \in G_N} P(G_N | v^* = v)$$

The quantity $P(G_N|v^* = v)$ (the likelihood of a virus graph given a source) will be difficult to compute, but we will find that this can be approached by a combinatorial problem.

3.2 Regular trees

We begin with regular trees, that is, trees where every node has the same degree. A virus graph construction can be simply described by a permutation of the N nodes, which isn't unique. However, not all permutations of the nodes can represent a potential virus graph, due to the structure constraint of the virus graph. Those permutations allowing us to reproduce the virus graph are called *permitted permutations.*

$P(G_N|v^* = v)$ can then be computed exactly : it will be the sum of the probabilities of all permitted permutations. On a regular tree, these probabilities are all equal to P[2], with

$$P = \frac{1}{k}\frac{1}{k + (k-2)} \cdots \frac{1}{k + (N-2)(k-2)}$$

To understand this, consider the source node : it has k free edges. When it joins a neighbour, the total number of free edges become $k + (k - 2)$, and so on. This gives us the probability of any N nodes permitted permutation, given any source in G_N.

Now, all we have to do is to compute the number of permitted permutations, that is, the number of ways the virus can spread.

Definition 1. *$R(v, T)$ is the number of permitted permutations of nodes that result in a tree T and begin with node $v \in T$.*

As a consequence, the likelihood turns out to be proportional to $R(v, T)$, so we have

$$\hat{v} = \arg\max_{v \in G_N} P(G_N | v^* = v) = \arg\max_{v \in G_N} R(v, G_N)$$

Here, $R(v, G_N)$ is called the **rumor centrality** of node v. The node which maximizes it is called the **rumor center**.

3.3 Computing rumor centrality

This is a combinatorial problem.

[2] "This is because of the memoryless property of the virus spreading time between nodes and the constant degree of all nodes." [SZ10]

Definition 2. *$T^v_{v_j}$ is the number of nodes in the subtree rooted at node v_j , with node v as the source.*

The goal is to count the number of permitted permutations of the N nodes, describing a virus diffusion beginning at source v and compatible with the virus graph G_N. The k neighbors of v are v_1, ..., v_k.

We have N slots free to describe the permutation. The first node has to be the source v. Then the nodes whose subtree is rooted at v_1 will be placed in the slots, and there are $T^v_{v_1}$ such nodes. They can be ordered in $R(v_1, T^v_{v_1})$ different ways. There remains $N-1-T^v_{v_1}$ nodes. We continue with the second neighbor, and so on. Multiplying all these quantities gives :

$$R(v, G_N) = \binom{N-1}{T^v_{v_1}} \binom{N-1-T^v_{v_1}}{T^v_{v_2}} ... \binom{N-1-\sum_{i=1}^{k-1} T^v_{v_i}}{T^v_{v_k}} \prod_{i=1}^{k} R(v_i, T^v_{v_i})$$

Playing around with this big expression gives in the end a simple formula for $R(v, G_N)$:

$$R(v, G_N) = N! \prod_{u \in G_N} \frac{1}{T^v_u}$$

To do this computation, Shah and Zaman proposed a linear-time algorithm in [SZ10].

3.4 General trees

To this point, we have defined and computed the rumor centrality for regular trees. In this simple case, the rumor centrality is an exact ML virus source estimator.

In general trees, all permitted permutations may not have the same probability. However, the rumor centrality remains a good estimator.

3.5 General graphs

To apply the precedent properties on trees, we shall consider the spanning tree induced by the virus graph. As this could be difficult to know, a Breadth-First-Search tree is often used instead.

For more details, see [SZ10] (early paper) and [SZ11] (more mature paper).

4 Multiple sources

These papers did only considered the case of a single rumor source. What if there was multiple sources ? Is it still manageable ?

For more details, see [LTL13] and [CZY14].

5 Suspects

Suppose you have a set of suspects, that is, an a priori knowledge. See [DZT13] for this.

6 Limited observations

Previous approachs supposed that the graph and the infected nodes could all be observed. In practice, this tends to be difficult, especially if the graph is a big one, such as with social networks. Thus, it would be interesting to develop new approachs needing only a finite and small number of observers.

Pedro Pinto, from EPFL, tried to address this in [PTV12].

See [SMA12] and [PTV12]. For multiple observations see [WDZT14]

7 Related projects

Some projects might be mentioned here because they are closely related to epidemiology.

- STEM (Spatiotemporal Epidemiological Modeler): `http://www.eclipse.org/stem/`
- Epigrass : `http://sourceforge.net/projects/epigrass/`
- NetLogo : `http://ccl.northwestern.edu/netlogo/`

References

[CZY14] Zhen Chen, Kai Zhu, and Lei Ying. Detecting multiple information sources in networks under the sir model. In *Information Sciences and Systems (CISS), 2014 48th Annual Conference on*, pages 1–4, March 2014.

[DZT13] Wenxiang Dong, Wenyi Zhang, and Chee Wei Tan. Rooting out the rumor culprit from suspects. *CoRR*, abs/1301.6312, 2013.

[LTL13] Wuqiong Luo, Wee Peng Tay, and Mei Leng. Identifying infection sources and regions in large networks. *Trans. Sig. Proc.*, 61(11):2850–2865, June 2013.

[PTV12] Pedro C. Pinto, Patrick Thiran, and Martin Vetterli. Locating the source of diffusion in large-scale networks. *CoRR*, abs/1208.2534, 2012.

[SMA12] Eunsoo Seo, Prasant Mohapatra, and Tarek Abdelzaher. Identifying rumors and their sources in social networks, 2012.

[SZ10] Devavrat Shah and Tauhid Zaman. Detecting sources of computer viruses in networks: Theory and experiment. *SIGMETRICS Perform. Eval. Rev.*, 38(1):203–214, June 2010.

[SZ11] D. Shah and T. Zaman. Rumors in a network: Who's the culprit? *IEEE Trans. Inf. Theor.*, 57(8):5163–5181, August 2011.

[WDZT14] Zhaoxu Wang, Wenxiang Dong, Wenyi Zhang, and Chee Wei Tan. Rumor source detection with multiple observations: Fundamental limits and algorithms. *SIGMETRICS Perform. Eval. Rev.*, 42(1):1–13, June 2014.