ITMO UNIVERSITY

Методы поиска источника распространения информации в комплексных сетях

Лыткин Юрий

Национальный центр когнитивных разработок,
Университет ИТМО, Санкт-Петербург

25 февраля 2021 г.

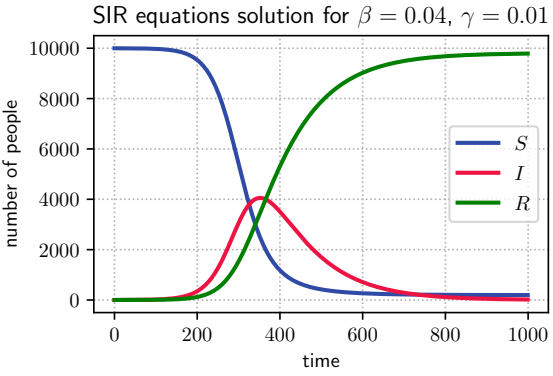
- Болезни и эпидемии.
- Компьютерные вирусы.
- Слухи в социальных сетях, fake news и вот это вот всё.

- Каждый человек в популяции находится в одном из трёх состояний:
  1. Susceptible, т.е. восприимчив к болезни, может заразиться.
  2. Infected, т.е. заражён и заражает других.
  3. Recovered, т.е. выздоровел, больше не подвержен болезни и не заражает других.
- Параметры:
  - $\beta \in [0, 1]$ отвечает за скорость заражения,
  - $\gamma \in [0, 1]$ — за скорость выздоровления.

- Пусть $S(t)$, $I(t)$, $R(t)$ — число, соответственно, восприимчивых, заражённых и выздоровевших в момент t.

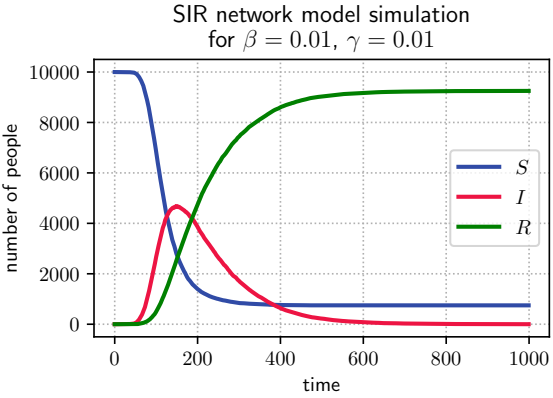
- Модель SIR :

$$\begin{cases} \dfrac{dS}{dt} = -\beta \dfrac{SI}{N} \\ \dfrac{dI}{dt} = \beta \dfrac{SI}{N} - \gamma I \\ \dfrac{dR}{dt} = \gamma I \end{cases}$$



Модель SIR на сети

- Допустим, у нас структура взаимодействий между участниками процесса (в виде сети).
- Модель SIR на сети. В течение каждого временного шага заражённая вершина может:
  - заразить каждого своего соседа с вероятностью β,
  - выздороветь с вероятностью γ.



Обратная задача. Поиск источника

Если задана сетевая структура под распространением, то можно сформулировать обратную задачу.

Обратная задача

Используя информацию о состоянии сети в некоторый момент времени, отыскать источник распространения.

Упрощения и предположения:

- Полностью известна структура сети. Сеть невзвешенная, неориентированная, без петель и кратных рёбер.
- Модель SI (заражённые вершины не выздоравливают), один источник распространения.

Формальная постановка

- $G = (V, E)$ — неориентированный граф без петель и кратных рёбер.
- В момент времени t_0 по графу G начинает распространяться сигнал, начиная с некоторой вершины-источника $s^* \in V$.
- Всякая вершина, получившая сигнал, начинает распространять его дальше.
- Если вершина u заражена, то за один временной шаг она заражает каждого своего соседа с вероятностью β.
- Вершины-наблюдатели $O = \{o_1, \ldots, o_k\} \subseteq V$ фиксируют время получения сигнала $T = \{t_1, \ldots, t_k\}$.
- Задача: найти источник по информации, полученной от наблюдателей.

- P into- T hiran- V etterli A lgorithm.

Pinto, Thiran, Vetterli. Locating the Source of Diffusion in Large-Scale Networks // Physical Review Letters, 2012

- Использует принцип максимального правдоподобия для поиска источника:

$$s = \underset{v \in V}{\operatorname{argmax}} \left[P\left(T \mid s^* = v\right) \right],$$

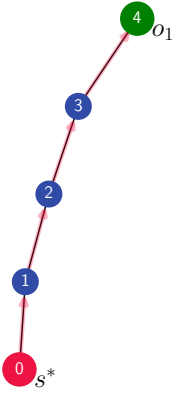
т.е. максимизируем вероятность иметь такие данные от наблюдателей при условии, что выбранная вершина является источником.

- Как посчитать вероятность $P(T \mid s^* = v)$?
- Сначала допустим, что G — дерево (т.е. не имеет циклов).
- Пусть o_i — i-й наблюдатель, t_i — время получения им сигнала. Значение t_i является реализацией случайной величины

$$\tau_i = t_0 + \sum_{(u,v) \in \rho(s^*, o_i)} \theta_{uv},$$

где

  - t_0 — время начала распространения (неизвестно),
  - $\rho(a, b)$ — список рёбер, составляющих кратчайший путь между a и b,
  - θ_{ab} — случайная величина, равная времени распространения сигнала по ребру $(a, b) \in E$ (имеет геометрическое распределение с параметром β).



- τ_i — задержка i-го наблюдателя.

- Чтобы избавиться от t_0, введём относительную задержку i-го наблюдателя:

$$\tau_i - \tau_1 = \sum_{(u,v)\in\rho(s^*,o_i)} \theta_{uv} - \sum_{(a,b)\in\rho(s^*,o_1)} \theta_{ab}, \;\; i \geq 2$$

- Из таких штук составим вектор относительных задержек:

$$\tau = (\tau_2 - \tau_1, \ldots, \tau_k - \tau_1)^\top$$

размера $k-1$, где k — число наблюдателей.

- Так и получается искомая вероятность:

$$P(T \mid s^* = v) = P(\tau = t \mid s^* = v),$$

где t — эмпирический вектор относительных задержек:

$$t = (t_2 - t_1, \ldots, t_k - t_1)^\top$$

- Осталось лишь посчитать эту вероятность. Дело за малым! (нет)
- Первая проблема: арифметические операции с геометрическим распределением — кошмар. Поэтому пока что «подменим» геометрическое распределение на нормальное.

Замечание

Здесь мы опираемся на Центральную предельную теорему: если путь между s^* и o_i достаточно длинный, то сумма $\sum_{(u,v)\in\rho(s^*,o_i)} \theta_{uv}$ имеет распределение, близкое к нормальному.

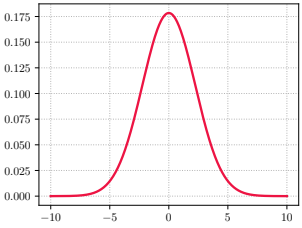
- Вторая проблема: нужно знать параметры распределения θ_{uv}. Геометрическое распределение с параметром β соответствует нормальному распределению с параметрами:

$$\mu = 1/\beta,\ \ \sigma = \sqrt{1-\beta}/\beta$$

- Итак, τ — $(k-1)$-мерный нормально распределённый случайный вектор. Как они вообще устроены?
- Нормальное распределение:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

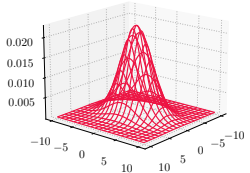
где μ — математическое ожидание, σ^2 — дисперсия.



- n-мерное нормальное распределение:

$$f(x) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \cdot \exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right),$$

где x — n-мерный вектор переменных, μ — вектор математических ожиданий, Σ — матрица ковариаций.



- В нашем случае это:

$$f_\tau(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_\tau|}} \cdot \exp\left(-\frac{1}{2}(x-\mu_\tau)^\top \Sigma_\tau^{-1}(x-\mu_\tau)\right)$$

  с параметрами μ_τ (вектор) и Σ_τ (матрица), которые зависят от положения предполагаемой вершины-источника в сети.
- Итак, принцип максимального правдоподобия для конкретной задачи:

$$s = \operatorname*{argmax}_{v \in V} \left[f_\tau(t \mid s^* = v) \right],$$

  где t — эмпирический вектор относительных задержек.
- Параметры μ_τ и Σ_τ можно посчитать, используя топологию сети. Сложно и долго, но возможно.

- В нашем случае это:

$$f_\tau(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_\tau|}} \cdot \exp\left(-\frac{1}{2}(x-\mu_\tau)^\top \Sigma_\tau^{-1}(x-\mu_\tau)\right)$$

  с параметрами μ_τ (вектор) и Σ_τ (матрица), которые зависят от положения предполагаемой вершины-источника в сети.
- Итак, принцип максимального правдоподобия для конкретной задачи:

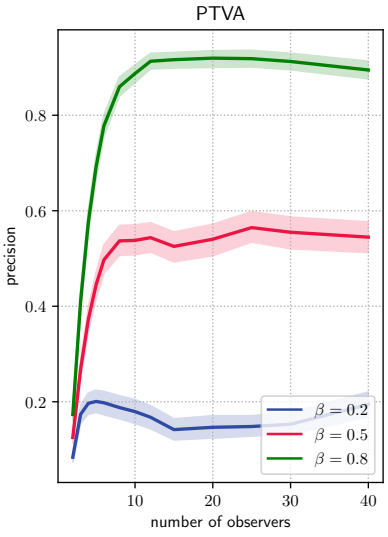
$$s = \underset{v \in V}{\operatorname{argmax}} \left[f_\tau(t \mid s^* = v) \right],$$

  где t — эмпирический вектор относительных задержек.
- Параметры μ_τ и Σ_τ можно посчитать, используя топологию сети. Сложно и долго, но возможно.
- Погодите, так это же всё работает только деревьев, нет? — Да, кстати, справедливо. Поэтому сначала нужно ~~посадить~~ построить дерево. PTVA использует ещё одно предположение: сигнал распространяется по дереву обхода в ширину.

PTVA. Итого

- Итак, алгоритм PTVA :
  1. Для данной вершины v строим дерево обхода в ширину.
  2. На основании этого дерева вычисляем вектор μ_τ и матрицу Σ_τ.
  3. Вычисляем $f_\tau(t)$, складываем его в общий список.

  Затем берём из общего списка вершину v с максимальным значением $f_\tau(t)$.

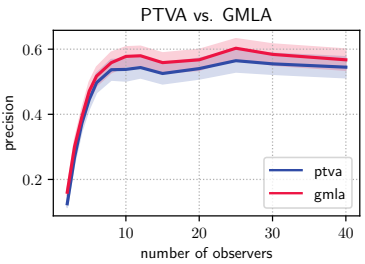


1. Сложность. Надо построить дерево обхода, по нему построить несколько матриц, да ещё и сделать всё это для каждой вершины $v \in V$. Имеем что-то вроде $O\left(n^3\right)$.
2. Предположение о том, что сигнал распространяется по дереву обхода в ширину, не всегда верно. Например, что если кратчайших путей несколько? Это влияет на значения ковариаций Σ_τ и приводит к ошибкам метода.

- В плане сложности на помощь приходит алгоритм GMLA (G radient M aximum L ikelihood A lgorithm).

Paluch, Lu, Suchecki, Szymanski, Hołyst. Fast and accurate detection of spread source in large complex networks // Scientific Reports, 2018

- Идейно похож на градиентный спуск:
  1. Берём какую-то вершину v, считаем для неё $f_\tau(t)$ (как в PTVA).
  2. Считаем $f_\tau(t)$ для каждого соседа v. Если где-то значение больше — идём туда. Если везде меньше (локальный максимум) — останавливаемся.



- Существуют и методы, учитывающие возможность распространения по одному из нескольких путей — EPP (E qui p robable P aths)
  и EPL (E qui p robable L inks).

Gajewski, Suchecki, Hołyst. Multiple propagation paths enhance locating the source of diffusion in complex networks // Physica A, 2019

- В них предложены более точные (и сложные) подходы к вычислению матрицы ковариаций Σ_τ.

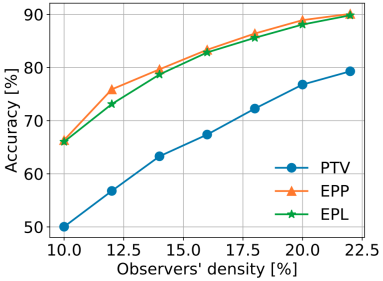


Рис.: Из работы авторов

Эвристические методы

- Методы, основанные на PTVA, являются точными в том смысле, что они точным образом (насколько это возможно) отражают поведение задержек наблюдателей. (Правда, только на деревьях...)
- Есть и другие, эвристические методы, которые не имеют в основе точный математический аппарат, а работают на основании каких-то иных предположений.

- T ime R eversal B ackward S preading.

Shen, Cao, Wang, Di, Stanley. Locating the source of diffusion in complex networks by time-reversal backward spreading // Phys. Rev. E, 2016

- Идея: рассмотрим массив значений

$$TR = \{t_i - \mu \cdot |\rho(s^*, o_i)|,\ i \geq 1\}$$

Если s^* — источник, то все эти значения должны быть *примерно* равны.

- TRBS:

$$s = \underset{v \in V}{\operatorname{argmin}} \left[\operatorname{Var}(TR \mid s^* = v) \right]$$

- Плюсы по сравнению с предыдущими алгоритмами:
  1. не требуются пути и их пересечения, а лишь расстояния
  2. не нужно знать параметр σ

- Авторы не дали этому методу название, так что просто Corr. Alg.

Xu, Teng, Zhou, Peng, Zhang, Zhang. Identifying the diffusion source in complex networks with limited observers // Physica A, 2019

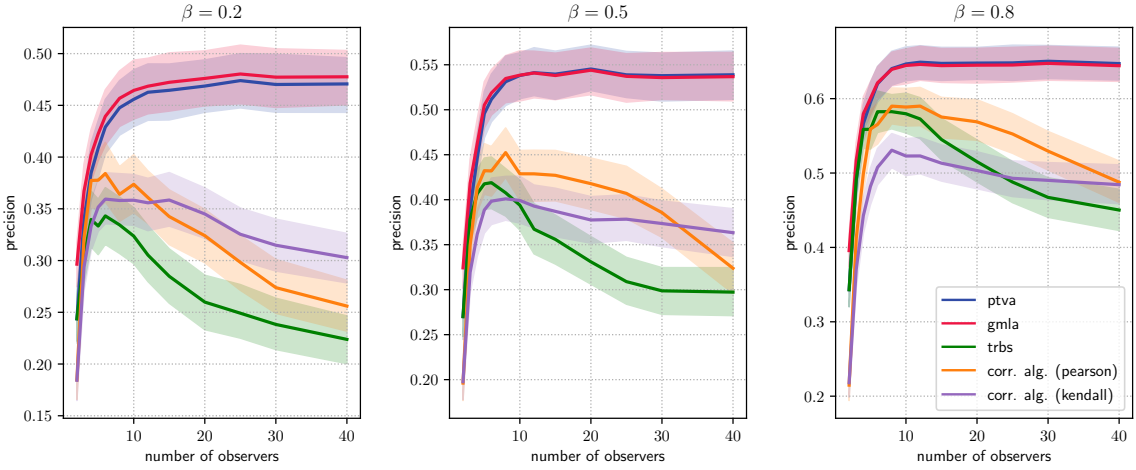
- Идея: если s^* — источник, между значениями $T = \{t_i,\ i \geq 1\}$ и $D = \{|\rho\,(s^*, o_i)|\,,\ i \geq 1\}$ должна быть высокая положительная корреляция.
- Corr. Alg. :

$$s = \underset{v \in V}{\operatorname{argmax}} \left[\operatorname{corr}(T, D \mid s^* = v) \right],$$

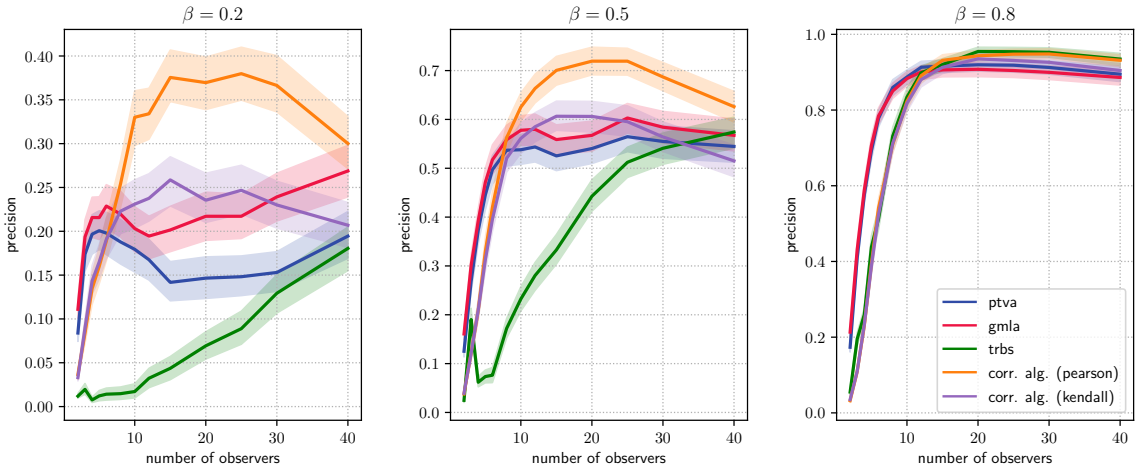
где corr — функция, реализующая коэффициент корреляции (например, коэффициент Пирсона или коэффициент Кендалла).

- Плюсы по сравнению с предыдущими алгоритмами:
  1. вообще никакие параметры не нужно знать

Сравнение методов. Случайные деревья



Сравнение методов. Графы Эрдёша-Реньи

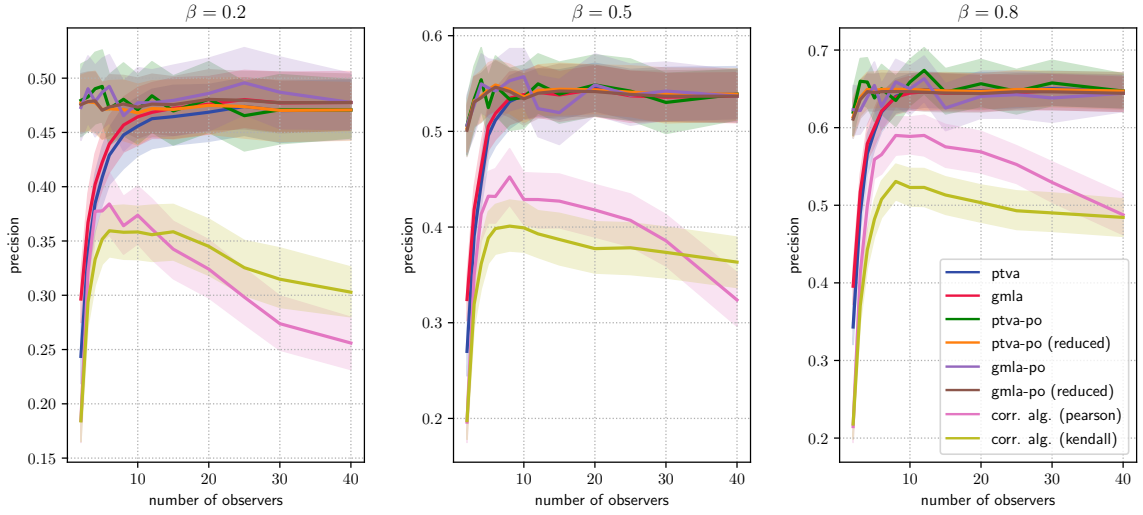


Пассивные наблюдатели

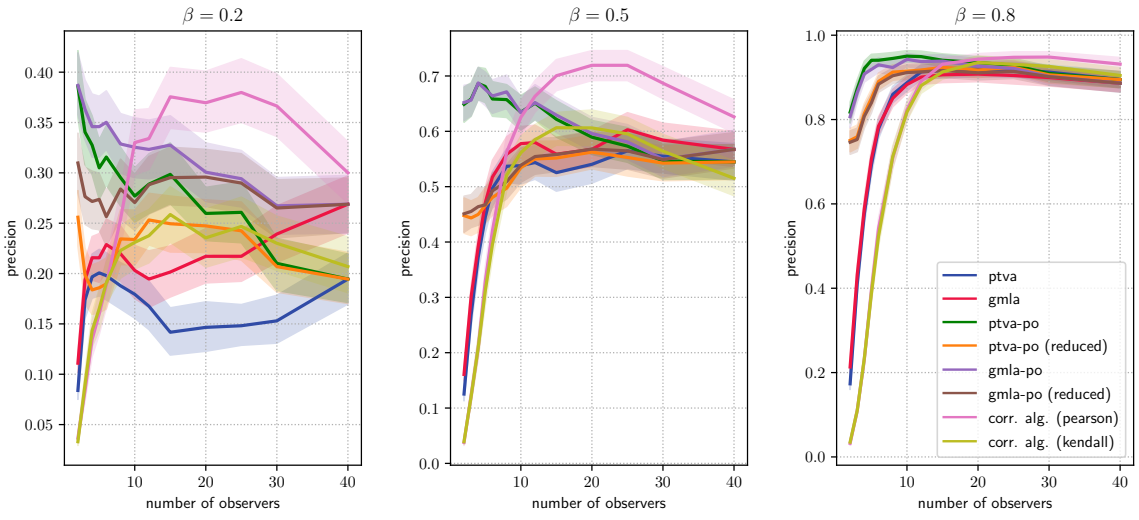
- Описанные ранее методы используют информацию от активных наблюдателей, т.е. тех, которые получили сигнал и зафиксировали время.
- А что если есть пассивные наблюдатели, которые никак не обозначили получение сигнала, поскольку на момент рассмотрения ещё его не получили?

- Идея: если у наблюдателя o_i нет значения t_i, то оно было бы больше, чем наибольшее t_j из имеющихся.
- В одномерном случае такую информацию можно использовать, переходя от вероятности $P(X = t)$ к вероятности $P(X > t)$. Для этого нужно проинтегрировать функцию плотности от t до ∞.
- В общем случае нужно проинтегрировать многомерную функцию плотности $f_\tau(x)$ по некоторым переменным от $t = \max\limits_i t_i$ до бесконечности.
- Правда, это аналитически невозможно.
- Но возможно численно!

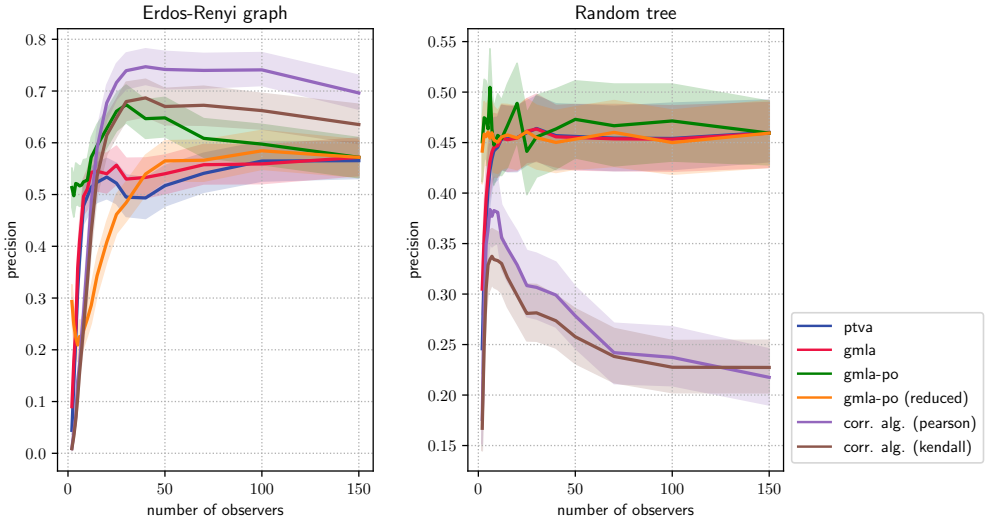
Сравнение методов. Случайные деревья



Сравнение методов. Графы Эрдёша-Реньи



Сравнение методов. Большие графы



На будущее

- Точные методы (типа PTVA) можно классно улучшать и развивать.
- Но они очень медленные и слишком много просят.
- Перспективная задача на будущее — развитие других (эвристических) методов на случай пассивных наблюдателей.