Correcting for Missing Data in Information Cascades

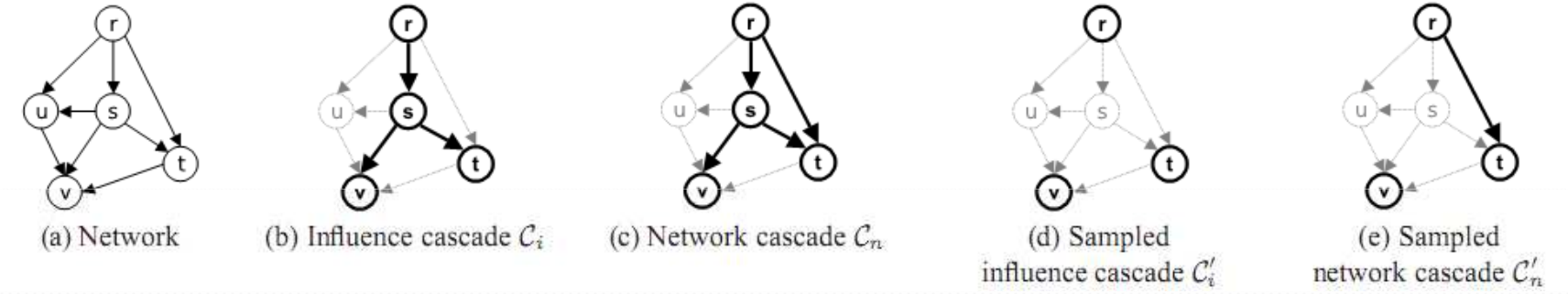
E. Sadikov, M. Medina,J Leskovec,H. Garcia-Molina
Stanford Universit

Докладчик Е. Моренко

План презентации

- Знакомство с каскадами
- Постановка задачи
- Методология
- K-деревья
- Оценки параметров k-деревьев
- Оценки модели
- Эксперименты
- Результаты

Знакомство с каскадами



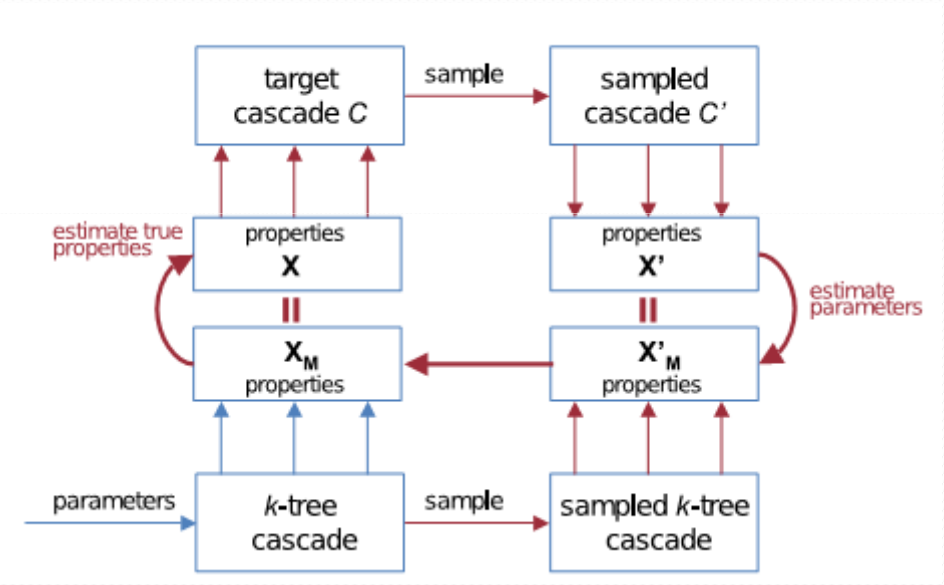
(a) Network (b) Influence cascade $\mathcal{C}_i$ (c) Network cascade $\mathcal{C}_n$ (d) Sampled influence cascade $\mathcal{C}_i'$ (e) Sampled network cascade $\mathcal{C}_n'$

Каскад C={($\perp$, r),(r,s)...}

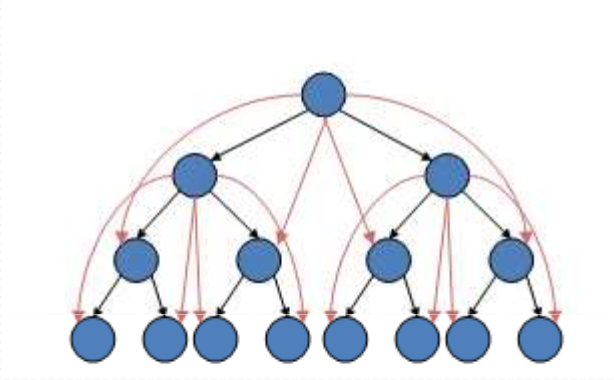
Постановка задачи

- Рассматриваем отсутствующие данные, как каскад C`, в котором вершины из C появляются с вероятностью σ
- Задача получить оценки для исходного каскада C.

Методология.



K-деревья



h=3, b=2, k=2

K-дерево Г(b,h,k) генерируется из сбалансированного дерева у которого глубина h, число потомков у каждого узла –b, число родителей k.

Простое k-дерево - Г(p,b,h,k) - k-дерево где 1-p – вероятность потери информации об узле.

Оценки параметров k-деревьев

Число узлов

Теорема 1.

Среднее число узлов в простом k-дереве - Г(p,b,h,k) :

$$m = p\frac{b^{h+1}-1}{b-1}$$

Оценки параметров k-деревьев

Число ребер

Теорема 2.

Среднее число ребер в простом k-дереве - Г(p,b,h,k) :

$$\frac{p^2}{b-1}\left(\frac{b(1-b^k)}{b-1}+kb^{h+1}\right)$$

Оценки параметров k-деревьев

Число изолированных вершин

Теорема 3.

Среднее число изолированных вершин в простом k-дереве - Г(p,b,h,k) :

$$\sum_{i=0}^{h} b^i p(1-p)^{l+\frac{b^{c+1}-b}{b-1}}$$

$$l = \min\{i,k\} \; and \; c = \min\{h-i,k\}.$$

Оценки параметров k-деревьев

Число слабо связанных вершин.

Теорема 4.

Среднее число слабо связанных вершин в простом k-дереве - Г(p,b,h,k) :

$$p\frac{[(1-p)b]^{a+1}-1}{(1-p)b-1}+\begin{cases} p(1-p)^k\frac{b^{h+1}-b^a}{b-1} & if \quad h>k \\ 0 & if \quad h\le k \end{cases}$$

$$a=\min(\{k,h\}).$$

Оценки параметров k-деревьев

Число листьев.

Теорема 5.

Среднее число листьев в простом k-дереве - Г(p,b,h,k) :

$$\sum_{i=0}^{h} b^i p(1-p)^{\frac{b^{c+1}-b}{b-1}}$$

$$c = \min\{h-i, k\}.$$

Теорема 6.

Среднее степень вершины, не являющийся листом в простом k-дереве - Г(p,b,h,k) при k=1 :

$$\frac{pb}{1-(1-p)^b}$$

Оценки параметров k-деревьев

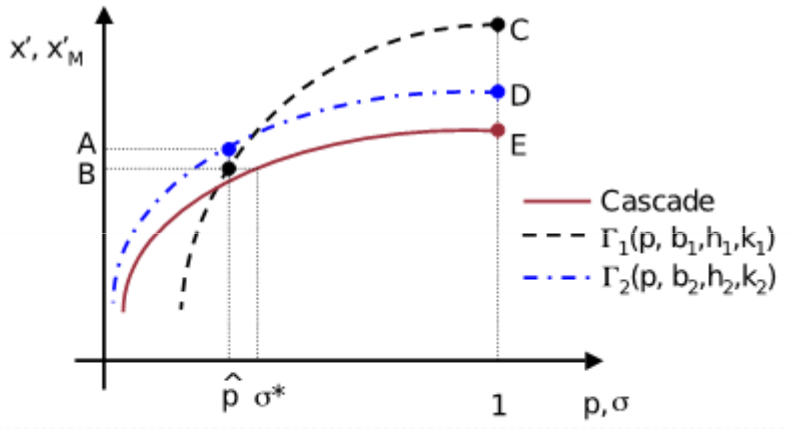
Средняя степень вершины.

Теорема 7.

Средняя степень вершины в простом k-дереве - Г(p,b,h,k) при h>>k оценивается как pk

Оценки модели

•Если известна σ, то оценку для p можно взять p=σ.
•Если σ неизвестна, то мы должны оценить σ отталкиваясь от природы данных.



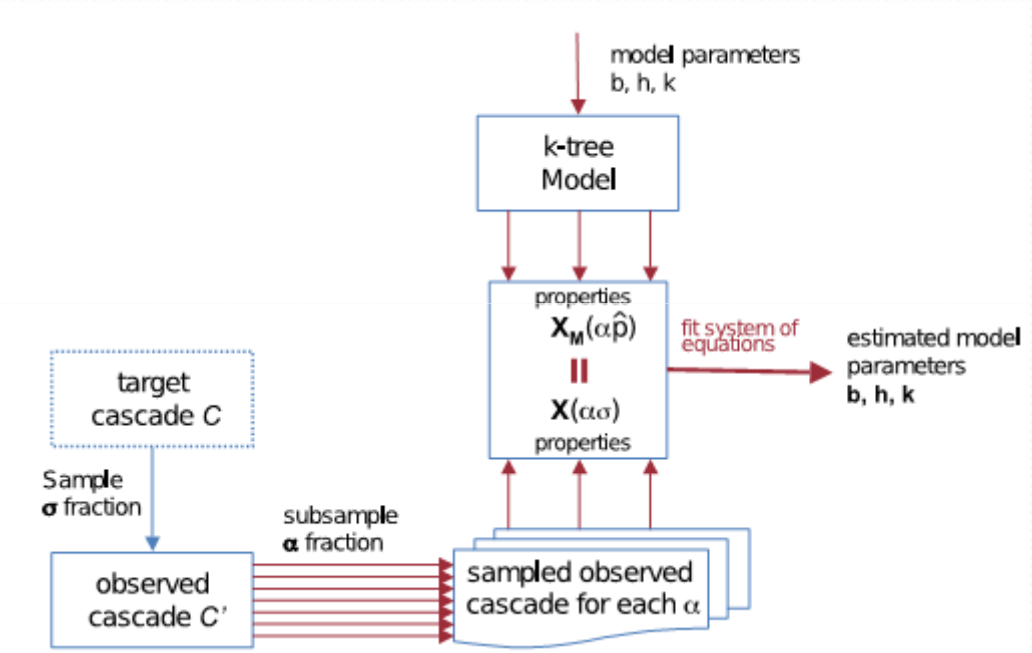
Подставляя полученную оценку для p в формулы из теорем 1-5 и вычисляя параметры C`, получим систему уравнений с 3 неизвестными b,h,k.

Данная система не линейна и мы не можем прямо её разрешить. Решая приближенно и минимизируя квадрат ошибки, авторы установили, что данный подход приводит к плохим результатам.

Оценки модели

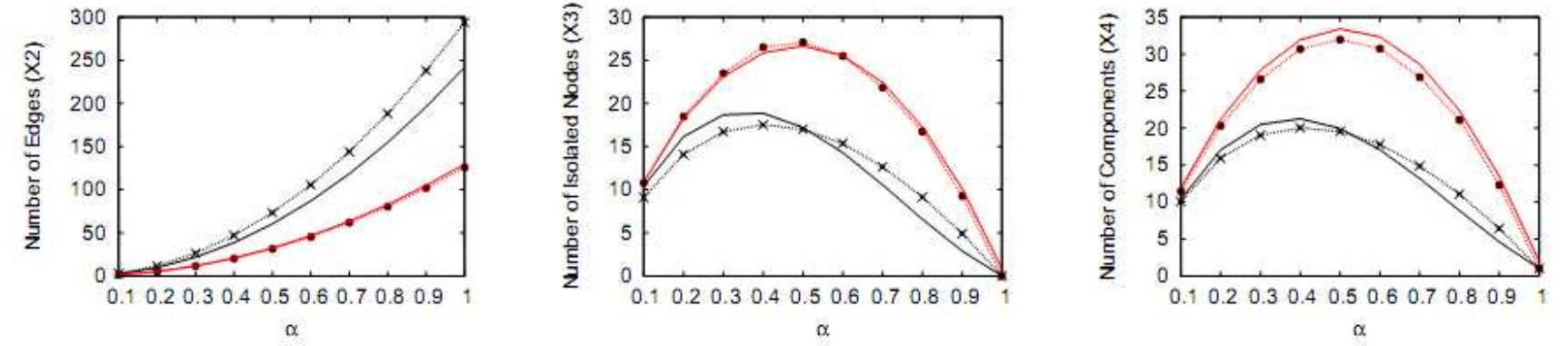
Основная идея - повысить точность, увеличивая количество точек для минимизации квадрата ошибок.

Для получения новых точек рассмотрим упрощение C` с параметром α (0< α ≤ 1). Получим простой каскад образованный из каскада C с параметром ασ.

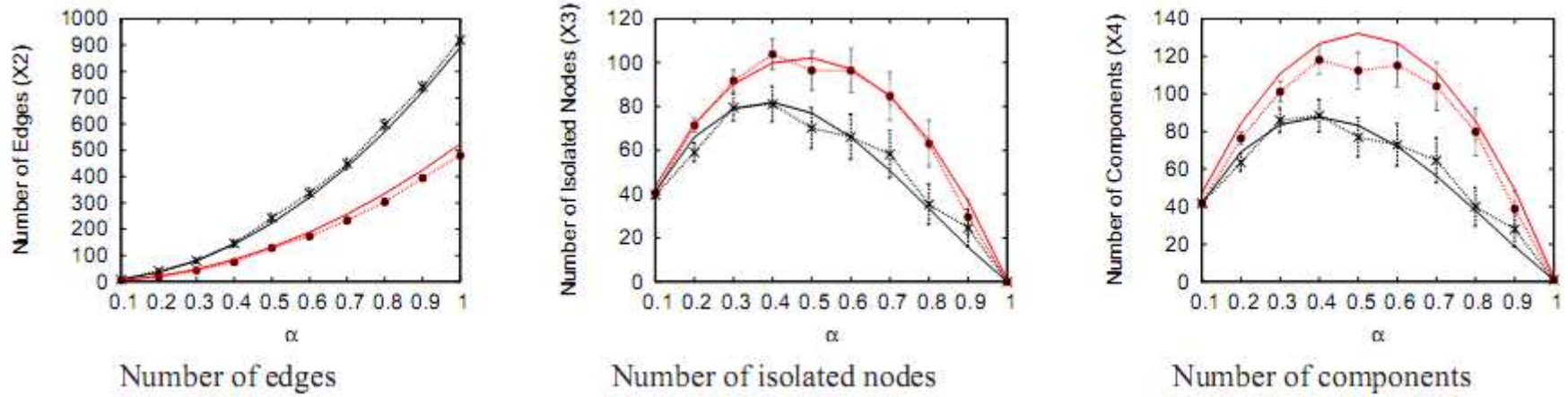


Что позволяет нам оценить квадрат ошибки на интервале (0, ασ).

Эксперименты

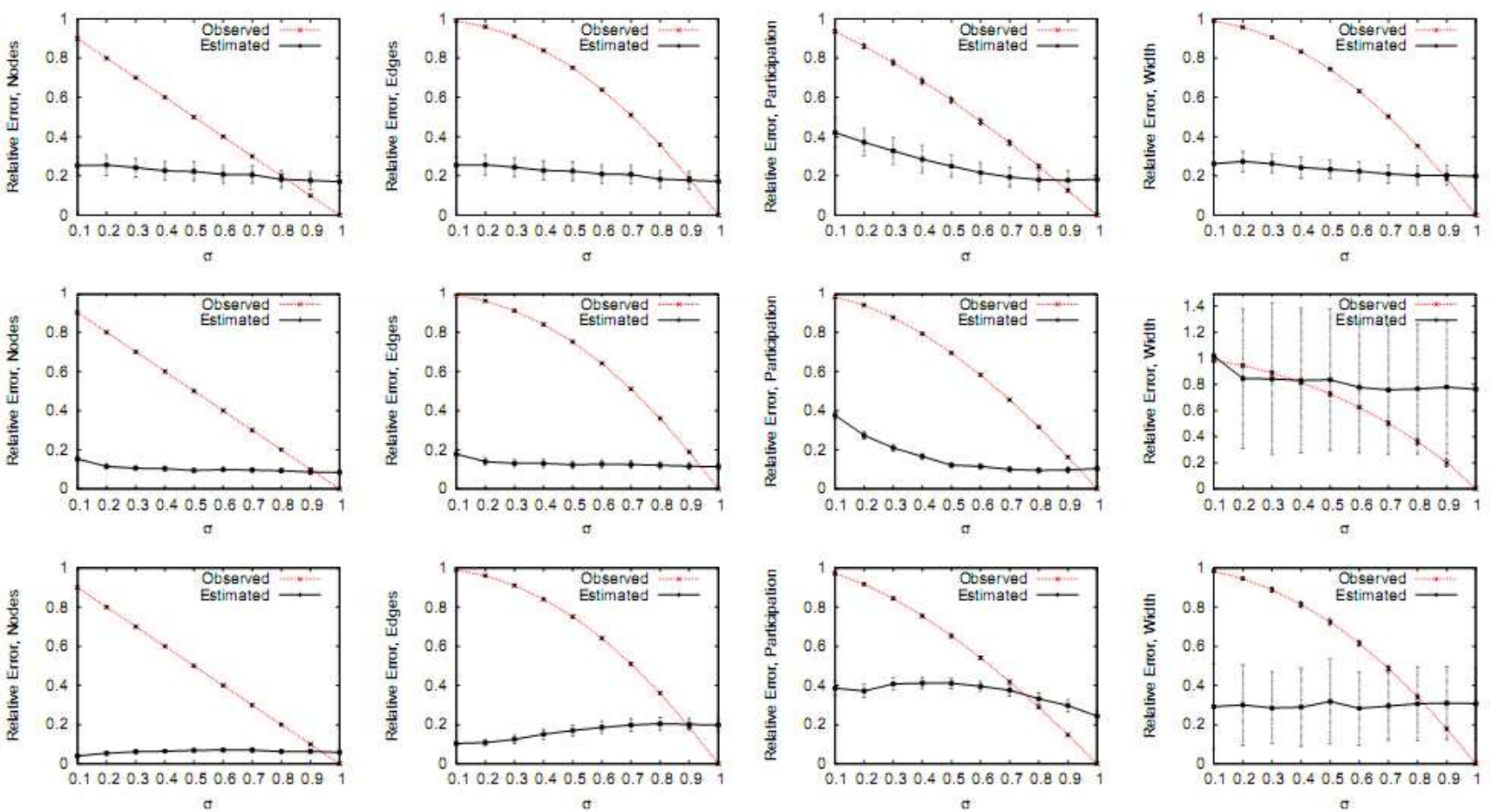


Оценки для каскадов основанных на твиттере.



Оценки для каскадов основанных на ретвитте.

Результаты


Observed
Estimated
Relative Error, Nodes
Relative Error, Edges
Relative Error, Participation
Relative Error, Width
α
Number of nodes
Number of edges
Participation
Cascade width

Спасибо за внимание.