Методы детектирования аномалий. Лекция 9: Аномалии в графах

Иван Шанин
ivan.shanin@gmail.com

ИПИ РАН

15.04.2019

Вершинные аномалии: egonet

Подход egonet заключается в выделении признаков из 1-окрестности каждой вершины, тем самым описывая каждую вершину набором признаков, например таких:

- (Вершинный признак n_i) Количество вершин в 1-окрестности вершины i (степень вершины)
- (Реберный признак e_i) Общее число ребер в 1-окрестности вершины i.
- (Весовой признак w_i) Для взвешенных графов — общий вес всех ребер в 1-окрестности вершины i.
- (Спектральный признак λ_i) Главное собственное значение взвешенного подграфа, состоящего из 1-окрестности вершины i.

Вершинные аномалии: egonet

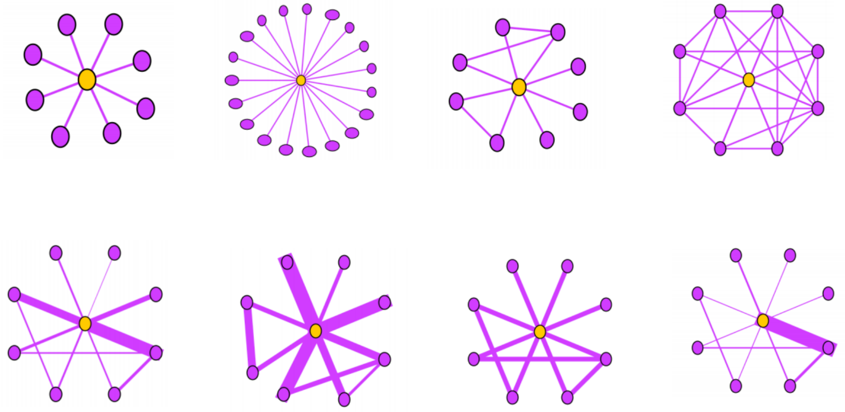

Рис. 1: различные состояния 1-окрестности вершины графа

Вершинные аномалии: egonet

Возможны следующие сценарии использования egonet (power laws):

- Вершинный признак vs реберный признак: таким образом можно отличить вершины, окрестность которых похожа на клику, от вершин, окрестность которых похожа на «звезду». Модель: $e_i \propto n_i^\alpha$, где $\alpha \in (1, 2)$, по графу оценивается параметр α, и вычисляется отклонение от соотношения.
- Весовой признак vs реберный признак. Аномальное состояние: «тяжелая» окрестность. Модель: $w_i \propto e_i^\beta$, где $\beta \geq 1$.
- Спектральный признак vs весовой признак: в этом сценарии аномалия определяется наличием тяжелого ребра в 1-окрестности. Модель: $\lambda_i \propto w_i^\gamma$, где $\gamma \in (0.5, 1)$

Аномалии связности

- Пусть A — это матрица смежности исследуемого графа, тогда представим

$$A \approx UV^T = \sum_{i=1}^{k} U_i, V_i^T$$

- U, V — матрицы размерности $n \times k$, где k - ранг факторизации
- Поиск U и V должен удовлетворять следующим условиям:

$$||A - UV^T||_F \to \min, \quad U, V \geq 0$$

- $R = A - UV^T$ — матрица остатков

Аномалии связности: матричные факторизации

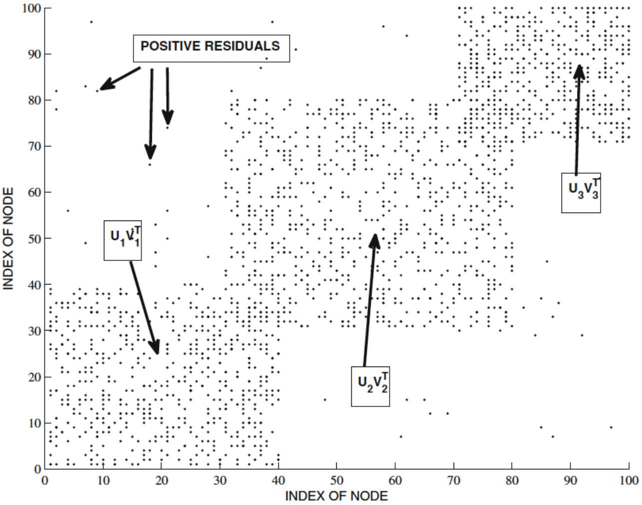

Рис. 2: Интерпретация NMF

Вероятностные графы

Построим оценку правдоподобия графа G

- $\mathcal{C} = \{C_1, C_2, \dots, C_k\}$: разбиение графа на компоненты
- Модель генерации ребер: k^2 вероятностей $p_{ij}(\mathcal{C})$ — вероятностей того, что случайно выбранное в графе ребро будет соединять компоненты i и j
- $\mathcal{F}(i, j, \mathcal{C})$: правдоподобие ребра (i, j) относительно разбиения $\mathcal{C}$
- Пусть есть r разбиений: $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_r$. Тогда определим усредненное правдоподобие $\mathcal{MF}(i, j, \mathcal{C}_1 \dots \mathcal{C}_r)$ ребра (i, j) как медианное значение $\mathcal{F}(i, j, \mathcal{C}_1) \dots \mathcal{F}(i, j, \mathcal{C}_r)$
- $\mathcal{GF}(G, \mathcal{C}_1 \dots \mathcal{C}_r)$ — правдоподобие графа G относительно разбиений $\mathcal{C}_1 \dots \mathcal{C}_r$

$$\mathcal{GF}(G, \mathcal{C}_1 \dots \mathcal{C}_r) = \left[\prod_{(i,j) \in G} \mathcal{MF}(i, j, \mathcal{C}_1 \dots \mathcal{C}_r) \right]^{\frac{1}{|G|}}$$

Аномальные подграфы: MDL

- $DL(G|S)$ — длина описания G при «известном» подграфе S
- У часто встречающихся подграфов будет низкий показатель F_1.

$$F_1(S,G) = DL(G|S) + DL(S)$$

- Низкий показатель F_2 является признаком аномальности подграфа

$$F_2(S,G) = Size(S) * Instances(S,G)$$

- Метод SUBDUE - итерационное «сжатие» графа:

$$O = 1 - \sum_{i=1}^{n} \frac{n-i+1}{n} \cdot \frac{DL_i - 1(S) - DL_i(S)}{DL_0(S)}$$