Сеть соавторства и методы вычисления вкладов соавторов

Щербакова Н. Г., Бредихин С.В., Ляпунов В.М.

Одним из классов комплексных сетей являются “сети сотрудничества”, в которых узлы связаны между собой членством в некоторой группе.

В нашем случае это группы авторов, совместно подготовивших общую научную публикацию (НП) – соавторы.

Проблемы:

- Роль, которую играет сотрудничество в научных исследованиях;
- Мотивации, побуждающие к сотрудничеству;
- Степень и уровни сотрудничества.

Индексы научного сотрудничества

Пример:

- Индекс сотрудничества: $CI = \frac{\sum_{r=1}^{n} r \cdot |P_r|}{|P|}$.

$CI \neq 0$, если $P \equiv P_1$

- Степень сотрудничества: $DC = \frac{|P_+|}{|P|}$ Значения на [0, 1]

Здесь:

P – множество всех публикаций;

P_+ – множество публикаций, имеющих более одного автора;

P_r – множество публикаций, имеющих r авторов.

Методы назначения вкладов авторов

Пусть n_k - число авторов публикации P_k.

1, 2, …, n_k - позиции в списке авторов; *w*(*i*) –вес автора *i.*

a. *w*(1) = 1 & *w*(*i*) = 0 , *i* = 2, …, n_k - все доверие первому автору, остальные имеют вес, равный нулю.

b. *w*(1), *w*(2), …, *w*(n_k) – веса убывают по мере увеличения номера позиции. Например:

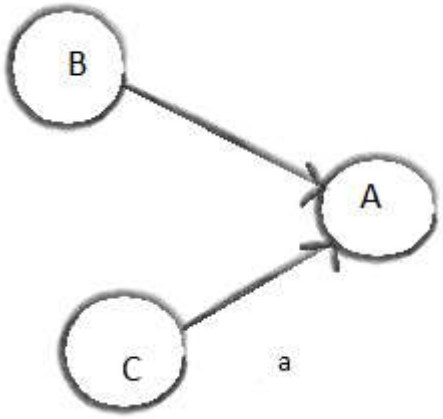
b.1. *w*(*i*) = 1/n_k с последующим перераспределением веса в сторону увеличения веса предыдущих авторов. При этом:

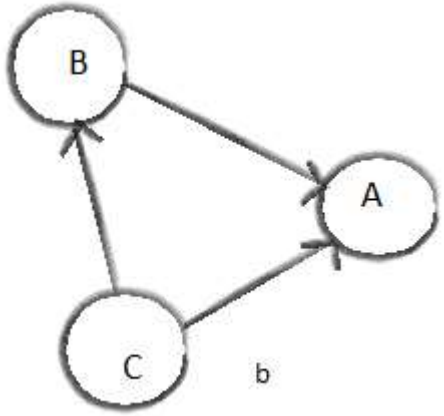
w(1) + *w*(2) +…+ *w*(n_k) = 1.

Методы назначения вкладов авторов

b.2. $w(i) = 1/i$ – вес обратен номеру позиции.

Ориентированные сети. Веса ребер зависят от числа совместных работ и позиций в списках авторов.





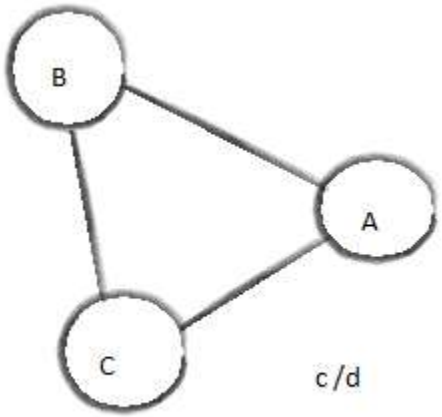
c. $w(i) = 1, i = 1, \ldots, n_k$ - полный вес (T).

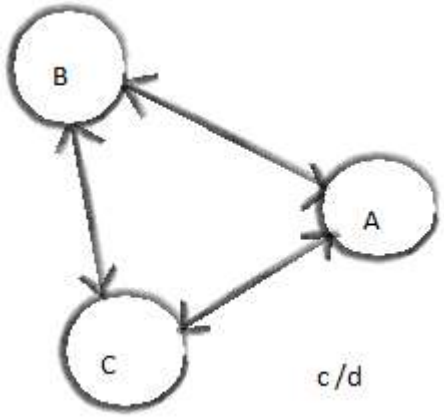
$w(1) + w(2) + \ldots + w(n_k) = n_k$.

d. $w(i) = 1/ n_k, i = 1, \ldots, n_k$ - частичный вес (F).

$w(1) + w(2) + \ldots + w(n_k) = 1$.

Неориентированные сети / ориентированные сети





Рассматриваемый сетевой подход

Пусть P – множество НП; V – множество авторов НП.

Определим две сети.

1. Сеть публикаций N^{pub} – это двудольная ориентированная сеть с узлами, разделенными на два непересекающихся независимых множества V и P, при этом дуга $e = (v_i, p_j) \in E^{pub}$, если v_i является автором / соавтором НП $p_j \in P$.
2. Сеть соавторства N^{ca} – это взвешенная неориентированная сеть $N^{ca} = (V, E, W)$, в которой V – множество авторов, $E \subseteq V \times V$ – множество взвешенных ребер, $e = (v_i, v_j) \in E$, если v_i *и* v_j являются соавторами хотя бы одной $p_k \in P$, W – матрица весов ребер.

Пусть |P| = l, |V| = n.

Пусть A - ($n \times l$) матрица авторства $A = (a_{ij})$, где $a_{ij} = 1$, если v_i является автором НП p_j , тогда ($n \times n$) матрица $U = A \times A^T$ с заменой всех ненулевых элементов на единицу является матрицей соавторства $U = (u_{ij})$:

$u_{ij} = \sum_{k=1}^{l} a_{ik}\, a_{jk}$, $u_{ij} = 1$, если v_i, v_j имеют хотя бы одну совместную НП.

Матрицу U можно рассматривать как (0,1) матрицу смежности сети соавторства и как матрицу смежности невзвешенной сети соавторства. Числа Эрдёша.

Определение ($n \times n$) матрицы весов $W = (w_{ij})$.

Т метод. Вес ребра w_{ij} - это число совместных работ авторов v_i и v_j. Каждая совместная работа добавляет 1 к весу ребра $e = (v_i, v_j)$.

Полный вес: $w_{ij} = u_{ij}$, где $U = A \times A^{\mathrm{T}}$ (с заменой $u_{ii} = 0$).

Нормированный полный вес: $w_{ij} = \frac{u_{ij}}{\sqrt{|P_i||P_j|}}$, где

P_i, P_j– множества публикаций авторов v_i, v_j соответственно.

F метод. Пусть n_k - число авторов публикации P_k. Каждая совместная публикация P_k добавляет $1/(n_k - 1)$ к значению веса ребра $e = (v_i, v_j)$, где v_i, v_j, - соавторы P_k. Т.е. вес ребра зависит от числа соавторов каждой отдельной НП, в которых участвовали v_i, v_j.

Частичный вес: $w_{ij}^* = \sum_{k=1}^{l} (a_{ik} \cdot a_{jk})/(n_k - 1)$,

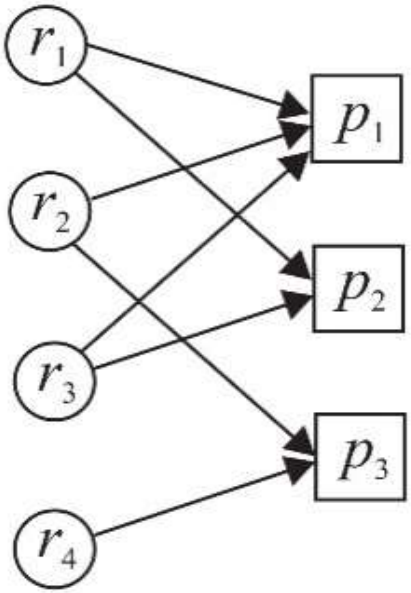
где $n_k = \sum_{i=1}^{n} a_{ik}$ число авторов НП k.

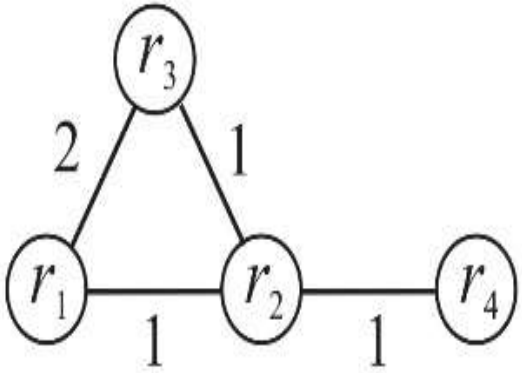
Пример.

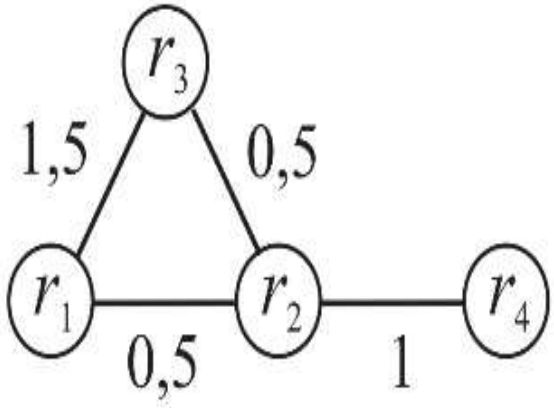
Фрагменты сети публикаций и двух сетей соавторства

$r_1 – r_4$ – авторы, $p_1 – p_3$ – публикации.

T- и F-методы расчета весов







Общий вес ребер, связывающий автора с соавторами при F-методе равен числу НП, опубликованных в соавторстве.

Разница между T- и F-методами подсчета весов ребер может оказаться несущественной при исследовании сетей соавторства, в которой шаблоны соавторства сходны.

В противном случае используемые подходы могут представить существенно различные картины сотрудничества. В частности, зависимость от метода определения весов может проявиться при вычислении мер центральности авторов.

Вычислительный эксперимент.

Анализ структуры и измерение параметров невзвешенной сети соавторства на основе информации, извлеченной из БД RePEc (2020.01.31).

Используются пакеты igraph, Pajek, Lapack.

Решение проблемы идентификации авторов на основе

“профилей”, создаваемых с помощью *Author Service*, БД RePEc.

Авторов – 32434.

Публикаций - 364979, из них 91113 имеют более одного автора.

Степень сотрудничества DC = 0,25

Размеры сети $N^{ca} = (V^{ca}, E^{ca})$: $n = |V^{ca}| = 32434$, $m = |E^{ca}| = 73969$. Размеры максимальной компоненты:

$N = (V, E)$: $|V| = 29270$ (~ 90%), $|E| = 71780$ (~ 97%).

Основные свойства:

- Распределение степеней узлов: $p_k \sim k^{-\gamma}$, $k_{min} = 4$, $\gamma = 1{,}3$. Среднее значение степени $\langle k \rangle = 4{,}6$.
- Характеристическая длина пути (среднее расстояние): $L(N^{ca}) \sim 6{,}58$; $L_{rand} \sim 6{,}74$.
- Локальный коэффициент кластеризации (для узла - доля связных соседей, для сети - среднее): $CC(N^{ca}) \sim 0{,}2644$; $C_{rand} \sim 0{,}0001$.

Выводы

В анализируемой коллекции НП доля командной работы невелика и составляет приблизительно 25%. При этом преобладающей тенденцией является наличие двух соавторов (77 % НП). Большинство авторов опосредованно связано друг с другом – максимальная компонента связности включает 90% авторов.

Сеть N^{ca} является масштабно-инвариантной (степенной закон распределения степеней) и относится к сетям «малого мира» ($L \sim L_{rand}$, $CC \gg C_{rand}$).