

КОНСТРУИРОВАНИЕ МАТРИЦЫ СМЕЖНОСТИ МУЛЬТИПЛЕКСНОЙ СЕТИ БАНКОВСКИХ ОПЕРАЦИЙ

Егоркин А.А.

Российский Государственный Социальный Университет, Москва, Россия
2-5@bk.ru

Аннотация. В работе приведено определение мультиплексной сети. Показаны подходы к построению такой сети для банковских транзакций, приведены критерии, на основании которых можно судить о корректности используемых предположений.

Ключевые слова: мультиплексные сети, матрица смежности, PageRank, центральность.

Введение

Мультиплексной сетью принято называть сеть, матрицу смежности которой (супра-матрица смежности) схематически можно представить в следующем виде [7]:

$$A = \begin{bmatrix} A^{[1]} & E & \dots & E & E \\ E & A^{[2]} & \dots & E & E \\ \dots & \dots & \dots & \dots & \dots \\ E & E & \dots & A^{[n-1]} & E \\ E & E & \dots & E & A^{[n]} \end{bmatrix}, \quad (1)$$

где:

$A^{[l]}$ – матрица смежности слоя l размером m ;

E – единичная матрица размером m ;

n – количество слоев в мультиплексной сети;

m – количество узлов в каждом слое.

Визуально мультиплексную сеть можно представить следующим образом [6]:

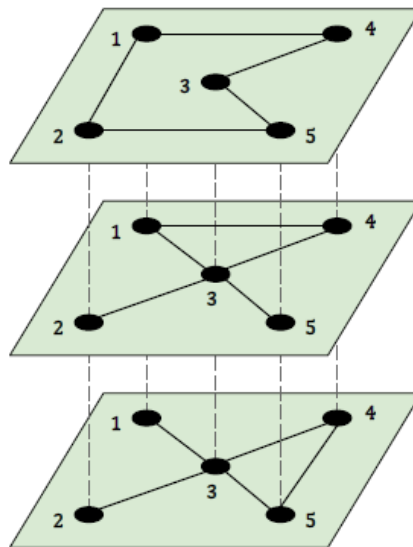


Рис. 1. Мультиплексная сеть для $n=3$ и $m=5$

Матрица A не является стохастической и не подходит для использования метрик центральности типа PageRank, а также не различает вероятность передвижения либо внутри слоя, либо по слоям.

Для того чтобы иметь возможность разделить «внутрислойное» и «межслойные» блуждания, а также чтобы иметь возможность применить алгоритм PageRank в [4] была предложена система коэффициентов, которая отражает вероятности указанных переходов. Однако, для большой сети количество данных коэффициентов растет нелинейно с увеличением количества узлов/слоев, что приводит к дискуссиям относительно того каким образом их выбирать [8]. В итоге авторы в ряде случаев использовали внешние константы, выбор которых мог серьезно отразиться на итоговом результате применения данной модели [5].

В настоящем докладе будут предложены наиболее простые подходы к построению матрицы смежности мультиплексной сети банковских транзакций, которые базируются на экономической природе используемых данных, что позволяет наиболее естественным образом получить итоговый результат с наименьшим количеством допущений.

1. Описание данных, используемых для построения мультиплексной сети

Мультиплексная сеть в настоящей работе будет построена на основании банковских операций. В качестве набора данных для построения мультиплексной сети используются все операции коммерческого банка в течение одного рабочего дня. Исходные данные представлены в следующей таблице 1:

Таблица 1. Исходные данные для анализа

Дата	Сумма платежа	Клиент отправителя платежа	Клиента получателя платежа	Счет клиента отправителя	Счет клиента получателя
DATE	SUM	SENDER_CLIENT	RECEIVER_CLIENT	SENDER_ACC	RECEIVER_ACC
		Тип клиента отправителя платежа	Тип клиента получателя платежа	Тип счета клиента отправителя	Тип счета клиента получателя
		SENDER_CLIENT_TYPE	RECEIVER_CLIENT_TYPE	SENDER_ACC_TYPE	RECEIVER_ACC_TYPE

Рассматривается два вида мультиплексной сети (G_1 и G_2), в первом случае на каждую из дат в используемом наборе данных строится мультиплексная сеть, где узлами являются клиенты, а слоями виды банковских операций. Во втором случае узлами выступают счета отправителей и получателей, а слоями взаимодействующие клиенты. Данные, представляющие собой клиентов плательщиков и получателей, состоят из одного перечня клиентов, т.к. один и тот же клиент может быть и плательщиком и получателем денежных средств в одну дату, аналогично и со счетами.

Для уменьшения количества узлов и слоев используем не конкретных контрагентов и счета, а типы(классы) клиентов и типы(классы) банковских счетов. Например, типы клиентов: физические лица, юридические лица, индивидуальные предприниматели и т.п. Типы счетов: депозитный счет, кредитный счет, текущий счет и т.п. Пример такой агрегации был рассмотрен на конференции [1].

Видами банковских операций для сети G_1 выступают всевозможные уникальные комбинации между типом счета отправителя и типом счета получателя. Видами клиентских взаимодействий для сети G_2 выступают всевозможные уникальные комбинации видов клиентов.

При построении слоев не учитывается их порядок, т.е. $SENDER_CLIENT_TYPE_1 \& SENDER_CLIENT_TYPE_2 = SENDER_CLIENT_TYPE_2 \& SENDER_CLIENT_TYPE_1$

Таблица 2. Компоненты мультиплексной сети

Параметры сети	Сеть G_1	Сеть G_2
Weight	SUM	SUM
Nodes	unique (SENDER_CLIENT_TYPE; RECEIVER_CLIENT_TYPE)	unique (SENDER_ACC_TYPE; RECEIVER_ACC_TYPE)
Layers	unique_sort (SENDER_ACC_TYPE & RECEIVER_ACC_TYPE)	unique_sort (SENDER_CLIENT_TYPE & RECEIVER_CLIENT_TYPE)

Также необходимо отметить, из-за того что при построении слоев используются всевозможные комбинации типов клиентов/счетов, то образуется достаточно большое количество слоев, содержащих малое количество связей. А учитывая, что финансовые операции подчинены степенному закону распределения [2], то все слои с суммарным весом связей ниже определенной величины целесообразно относить в технический слой – «Прочее».

2. Построение стохастической матрицы смежности мультиплексной сети

Для того чтобы сделать матрицу смежности мультиплексной сети стохастической, необходимо добиться ее стохастичности для каждого из слоев:

$$G^{[l]} = \alpha A^{[l]} + (1 - \alpha)J \quad (2)$$

$$\hat{G}_{i,j}^{[l]} = \begin{cases} G_{i,j}^{[l]}, & \text{if } \sum_{i=1}^m G_{i,j}^{[l]} = 1 \\ \frac{G_{i,j}^{[l]}}{\sum_{i=1}^m G_{i,j}^{[l]}}, & \text{if } \sum_{i=1}^m G_{i,j}^{[l]} < 1 \end{cases} \quad (3)$$

где:

$G^{[l]}$ – google матрица слоя l ;

$\hat{G}^{[l]}$ – стохастическая google матрица слоя l с заполненными висящими узлами;

α – коэффициент телепортации внутри слоя;

i, j – индексы матрицы смежности;

J – матрица единиц размером m .

Подробный порядок перехода на висящие узлы и подходы к определению коэффициента телепортации прописаны в [3].

Стохастическая матрица смежности мультиплексной сети определяется следующим образом:

$$\mathcal{A} = \delta \hat{\mathbf{G}} + (1 - \delta) \hat{\mathbf{E}} \otimes \mathbf{A} \quad (4)$$

$$\hat{\mathbf{G}} = \begin{bmatrix} \hat{G}^{[1]} & 0 & \dots & 0 & 0 \\ 0 & \hat{G}^{[2]} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \hat{G}^{[n-1]} & 0 \\ 0 & 0 & \dots & 0 & \hat{G}^{[n]} \end{bmatrix} \quad (5)$$

$$\hat{\mathbf{E}} = \begin{bmatrix} 0 & E & \dots & E & E \\ E & 0 & \dots & E & E \\ \dots & \dots & \dots & \dots & \dots \\ E & E & \dots & 0 & E \\ E & E & \dots & E & 0 \end{bmatrix} \quad (6)$$

$$\mathbf{A} = \begin{bmatrix} 0 & \lambda_{1,2} & \dots & \lambda_{1,n-1} & \lambda_{1,n} \\ \lambda_{2,1} & 0 & \dots & \lambda_{2,n-1} & \lambda_{2,n} \\ \dots & \dots & \dots & \dots & \dots \\ \lambda_{n-1,1} & \lambda_{n-1,2} & \dots & 0 & \lambda_{n-1,n} \\ \lambda_{n,1} & \lambda_{n,2} & \dots & \lambda_{n,n-1} & 0 \end{bmatrix} \quad (7)$$

где:

\mathcal{A} – стохастическая матрица смежности мультиплексной сети финансовых транзакций размерностью $n \cdot m$;

$\lambda_{l,k}$ – коэффициент телепортации со слоя l на слой k , в общем случае $\lambda_{l,k} \neq \lambda_{k,l}$;

δ – коэффициент глобальной телепортации: с вероятностью δ – осуществляется переход с одного узла на другой узел внутри слоя, с вероятностью $(1 - \delta)$ – переход происходит между слоями.

Основной задачей при конструировании матрицы \mathcal{A} является нахождение параметров $\lambda_{l,k}$ и δ . Как отмечалось выше в [5], авторы использовали в ряде случаев наиболее простые внешние константы при определении данных параметров, например:

$$\lambda_{l,k} = 1/(n - 1) \quad (8.1)$$

В настоящем исследовании будут использованы предположения, базирующиеся на физической природе рассматриваемых явлений.

Так будет показано, что определение коэффициентов телепортации между слоями, посчитанных исходя из веса слоя, будет давать результат лучше по сравнению с описанным выше подходом:

$$\lambda_{k,l} = \frac{\sum_{i,j} A_{i,j}^{[l]}}{\sum_{k,k \neq l} \sum_{i,j} A_{i,j}^{[k]}} \quad (8.2)$$

Параметр глобальной телепортации δ может быть определен методами машинного обучения.

3. Определение агрегированных показателей сети

Вектор центральности PageRank определяется следующим образом:

$$\lim_{t \rightarrow \infty} \mathbf{X}_t = \mathcal{A}\mathbf{X}_{t-1} \quad (9)$$

где:

\mathbf{X} – вектор центральности мультиплексной сети размером $n \cdot m$, имеет блочную структуру:

$$\mathbf{X} = [x_1^{[1]} \dots x_m^{[1]}, x_1^{[2]} \dots x_m^{[2]}, \dots, x_1^{[n-1]} \dots x_m^{[n-1]}, x_1^{[n]} \dots x_m^{[n]}]^T \quad (10)$$

где:

$x_i^{[l]}$ – центральность узла i в слое l .

Для целей дальнейшего исследования введем агрегированные показатели: среднюю центральность узла:

$$\bar{c}_i = \sum_{j=1}^{nm} \begin{cases} x_j & \text{if } (j \bmod n) = i \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

и центральность слоя:

$$\bar{c}^{[l]} = \sum_{j=1}^{nm} \begin{cases} x_j & \text{if } (j \div n) = l \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

4. Сравнение между собой агрегированных показателей сетей \mathcal{G}_1 и \mathcal{G}_2

Исходные данные интерпретируются следующим образом: в сети \mathcal{G}_1 слоями являются узлы (конкатенация узлов) сети \mathcal{G}_2 , а конкатенация узлов сети \mathcal{G}_1 является слоями сети \mathcal{G}_2 . Т.е. можно предположить, что центральность слоя \mathcal{G}_1 может определяться не только как сумма центральностей узлов, принадлежащих данному слою (12), но и как сумма компонентов, из конкатенации которых состоит слой:

$$\begin{aligned} \bar{c}_1^{[l]} &= \bar{c}_1^{[l_1]} + \bar{c}_1^{[l_2]}, \\ l &= l_1 \& l_2, \bar{c}_1^{[l_1]} > 0, \bar{c}_1^{[l_2]} > 0 \end{aligned} \quad (13)$$

где:

$\bar{c}_1^{[l_1]}$ – центральность первой компоненты слоя l в сети \mathcal{G}_1 ,

$\bar{c}_1^{[l_2]}$ – центральность второй компоненты слоя l в сети \mathcal{G}_1

С другой стороны для сети \mathcal{G}_2 центральность узла – это есть центральность компоненты слоя в графе \mathcal{G}_1 , т. к. узел \mathcal{G}_2 и компонента слоя \mathcal{G}_1 представляют собой один и тот же тип бухгалтерского счета. Таким образом, мы можем сравнить центральность одних и тех же величин, но полученных различным путем.

В сети \mathcal{G}_1 количество слоев больше количества уникальных компонентов слоя. Поэтому система уравнений (14) для всех слоев и узлов сети \mathcal{G}_1 будет переопределённой, и ее решение может быть получено, исходя из минимизации ошибки:

$$\begin{aligned} C_1^{[L]} &= \mathbf{B} \times C_1^{[L_1 \& L_2]} \\ \left\| C_1^{[L]} - \mathbf{B} \times C_1^{[L_1 \& L_2]} \right\| &\rightarrow \min \end{aligned} \quad (14)$$

где:

$C_1^{[L]}$ – вектор центральности слоев \mathcal{G}_1 размером n_l – количество слоев в \mathcal{G}_1 ;

$C_1^{[L_1 \& L_2]}$ – вектор центральности компонентов слоев \mathcal{G}_1 размером m_2 – количество узлов в \mathcal{G}_2 ;

\mathbf{B} – матрица размера $[m_2 \times n_l]$

Полученный вектор $C_1^{[L_1 \& L_2]}$ нормируем исходя из условий, что все его компоненты положительные и их сумма равна 1.

На рис.2 приведено сравнение показателей центральности компонентов слоя \mathcal{G}_1 и узлов \mathcal{G}_2 . Видно, что данные имеют кластерную структуру, расположенную на диагонали. Каждая точка представляется собой один день, цвет – компоненты слоев \mathcal{G}_1 либо узлы \mathcal{G}_2 (что тождественно, т.к. они представляют собой один набор данных – тип счета плательщика/получателя). В расчете использовалось около 400

дней, и каждый кластер содержит такое количество точек. При абсолютном точном решении рассматриваемой задачи, все указанные кластеры должны были бы находиться на главной диагонали, но такого не происходит в виду погрешностей в вычислениях и исходных данных, а также за счет допущений, на основании которых были вычислены рассматриваемые параметры.

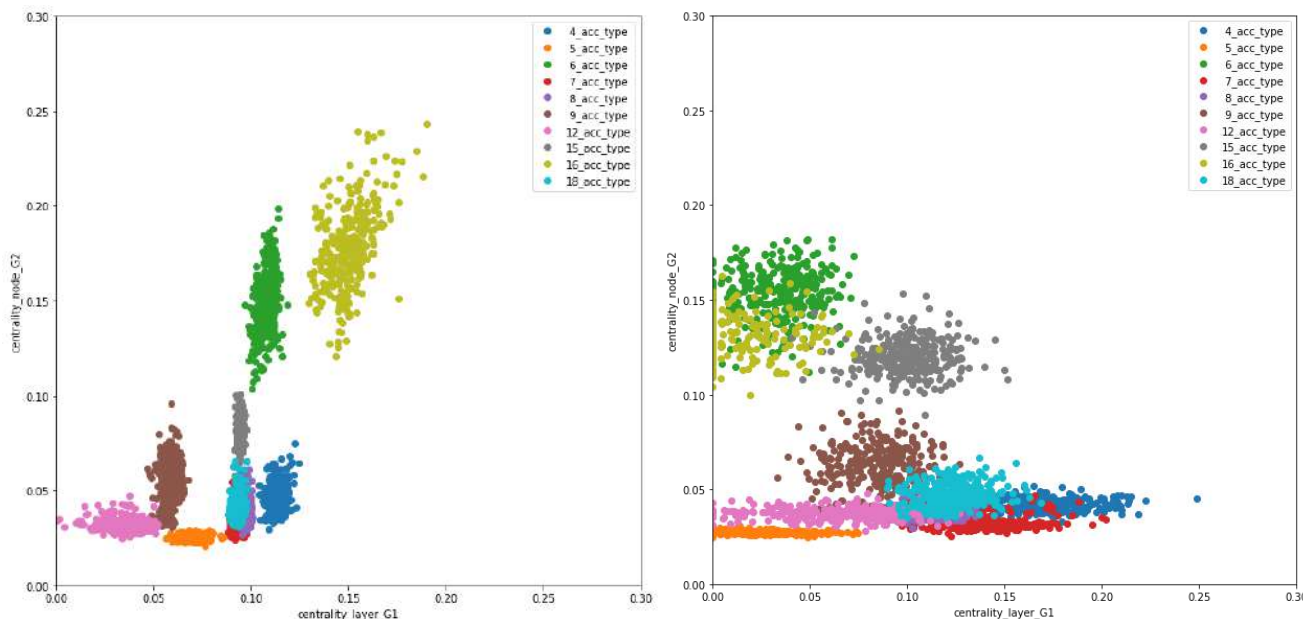


Рис. 2. Сравнение центральности компонентов слоя G_1 и узла G_2 для разных случаев коэффициентов $\lambda_{l,k}$, левый по формуле (8.2), правый по формуле (8.1)

На приведенных диаграммах рассеивания приведено сравнение центральностей двух сетей при различных коэффициентах $\lambda_{l,k}$. Как видно из приведенных данных использование формулы (8.2) дает лучший результат: кластеры плотные, сгруппированы по главной диагонали.

Количественным показателем качества полученных данных может выступать коэффициент корреляции Пирсона, который при использовании формулы (8.2) составляет более $2/3$, что свидетельствует о наличии зависимости между рассматриваемыми переменными, а при использовании формулы (8.1) имеет отрицательное значение.

5. Заключение

Предложенный подход к конструированию матрицы смежности мультиплексной сети позволяет использовать *большой* объем информации о сети банковских транзакций. Это позволит осуществлять моделирование финансовых процессов с увеличенной точностью, Например, улучшить качество прогнозирования краткосрочной ликвидности коммерческого банка, по сравнению с результатами, представленными в [1].

В работе предложены несколько вариантов конструирования матрицы \mathcal{A} , при этом они не являются единственными возможными, так в качестве слоев можно использовать типы операций, классифицированных на основании текстового разбора информации. Возможно использование *большого* количества параметров при построении матрицы смежности, например: валюты операции, срочности операции, риска операции и другие. Это позволит использовать не только мультиплексные сети, но и многослойные сети.

Литература

1. Егоркин А.А. Прогнозирование краткосрочной ликвидности коммерческого банка с использованием сетевых моделей // Управление развитием крупномасштабных систем (MLSD'2024): Труды Семнадцатой международной конференции, Москва, 24–26 сентября 2024 года. – Москва: Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В.А.Трапезникова Российской академии наук, 2024. – С. 600–606.
2. Егоркин А.А. Особенности использования алгоритма классификации k-means для данных, подчиненных степенному закону распределения // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. – 2023. – № 9. – С. 65–69.

3. *Егоркин А.А.* Определение центральности графа алгоритмом PageRank с учетом весов связей // Управление большими системами: сборник трудов. – 2024. – № 111. – С. 81–96.
4. *Baptista A., Gonzalez A. & Baudot A.* Universal multilayer network exploration by random walk with restart // Communications Physics. – 2022. – Vol. 5. – P. 170.
5. *Baptista A., Gonzalez A. & Baudot A.* Supplementary Information for Universal Multilayer Network Exploration by Random Walk with Restart
6. *Christian Ocklind.* Comparative analysis of multilayer network software, 202, P. 45
7. *Valdeolivas A, Tichit L, Navarro C, Perrin S, Odelin G, Levy N, Cau P, Remy E, Baudot A.* Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics. – 2019;35(3):497–505.
8. <https://www.nature.com/articles/s42005-022-00937-9#citeas>.