Междисциплинарный интеграционный проект СО РАН № 21

Исследование закономерностей и тенденций развития самоорганизующихся систем на примере веб-пространства и биологических сообществ

Блок 1. **Структурный, метрический и топологический анализ графов и сетей связей, возникающих в веб-пространстве, биологических и социальных сообществах**

ИМ, ИБФ, ИВТ, ИДСТУ, ИЦиГ

Согласно заявке:

Ожидаемые результаты:
– Характеристика изучаемых сетей, в том числе на основании их структурных и метрических инвариантов;
– Построение моделей организации и эволюции сетей различной природы.

Методы:
– Структурный анализ графов (цикличность структуры, симметрии, устойчивость к структурным изменениям);
– Кластерный анализ (включая оценки сложности кластерного анализа);
– Метрический анализ (глобальная и локальная характеризация на основе метрических и информационных инвариантов).

Структурные свойства сложных сетей:
– нагрузка вершины (число проходящих через нее кратчайших путей);
– подграфы (в частности, наличие клик);
– ассортативное или диссортативное перемешивание;
– кластеризация (транзитивность).

Типы сложных сетей:
– социальные;
– технологические;
– биологические.

Различие структурных свойств у сетей разных типов.

Кластеризация – выше у социальных сетей.

Ассортативность – положительная у социальных сетей, отрицательная у биологических сетей.

Newmann, 2002: расовые предпочтения при образовании супружеских пар в Сан-Франциско:

м ж	A	M	C	O
A	506	32	69	26
M	23	308	114	38
C	26	46	599	68
O	10	14	47	32

где A = Афроамериканцы, M = Мексиканцы, C = Светлокожие, O = Остальные.

Newmann, 2003: вычислен коэффициет Пирса r – показатель ассортативности (тяги к тому, чтобы связаться с вершиной той же степени) для многих сетей.

Социальные сети

Сеть	Тип	Размер	Ассортативность
соавторов по физике	неор.	52 909	0.363
соавторов по биологии	неор.	1 520 251	0.127
соавторов по математике	неор.	253 339	0.120
сотрудничества актеров кино	неор.	449 913	0.208
директоров компаний	неор.	7 673	0.276
связей студентов	неор.	573	–0.029 (?)
адресов электронной почты	ориент.	16 881	0.092

Технологические сети

Сеть	Тип	Размер	Ассортативность
сеть электростанций	неор.	4 941	–0.003
Интернет	неор.	10 697	–0.189
«Всемирная паутина» (WWW)	ориент.	269 504	–0.067
взаимозависимости программного обеспечения	ориент.	3 162	–0.016

Биологические сети

Сеть	Тип	Размер	Ассортативность
взаимодействий белков	неор.	2 115	–0.156
метаболическая сеть	неор.	765	–0.240
нейронная сеть	ориент.	307	–0.226
морская пищевая сеть	ориент.	134	–0.263
пресноводная пищевая сеть	ориент.	92	–0.326

Веб-графы

– ориентированный граф (возможно с кратными ребрами и кратными петлями): вершины – сайты; ребра – ссылки.

Barabashi – Albert, 1999: эмпирические свойства веб-графа:

– разреженность (если n вершин, то $m \cdot n$ ребер, $m =$const);

– диаметр графа равен от 5 до 7 (“теория 6 рукопожатий”, “мир тесен”);

– распределение вершин по числу связей в виде степенного закона:

$$\frac{\text{число вершин степени } = d}{\text{число всех вершин}} \sim C \cdot d^{-\gamma}, \qquad \gamma \approx 2.1$$

Рунет сегодня (?): 15 млн. сайтов, 200 млн. ссылок, $2.5 < \gamma < 3$.

Модели эволюции графа при присоединении новой вершины.

Модель Erdös – Renyi, 1959:

– теория случайных графов; биномиальное распределение:

$$P(\text{степень вершины } = d) = C_{n-1}^{d} \cdot p^{d}(1-p)^{n-1-d},$$

где n – число вершин графа, $m \cdot n$ – число ребер, $p = \frac{2m}{n} \in [0;1]$ – вероятность появления ребра между n-ой и i-ой вершиной.

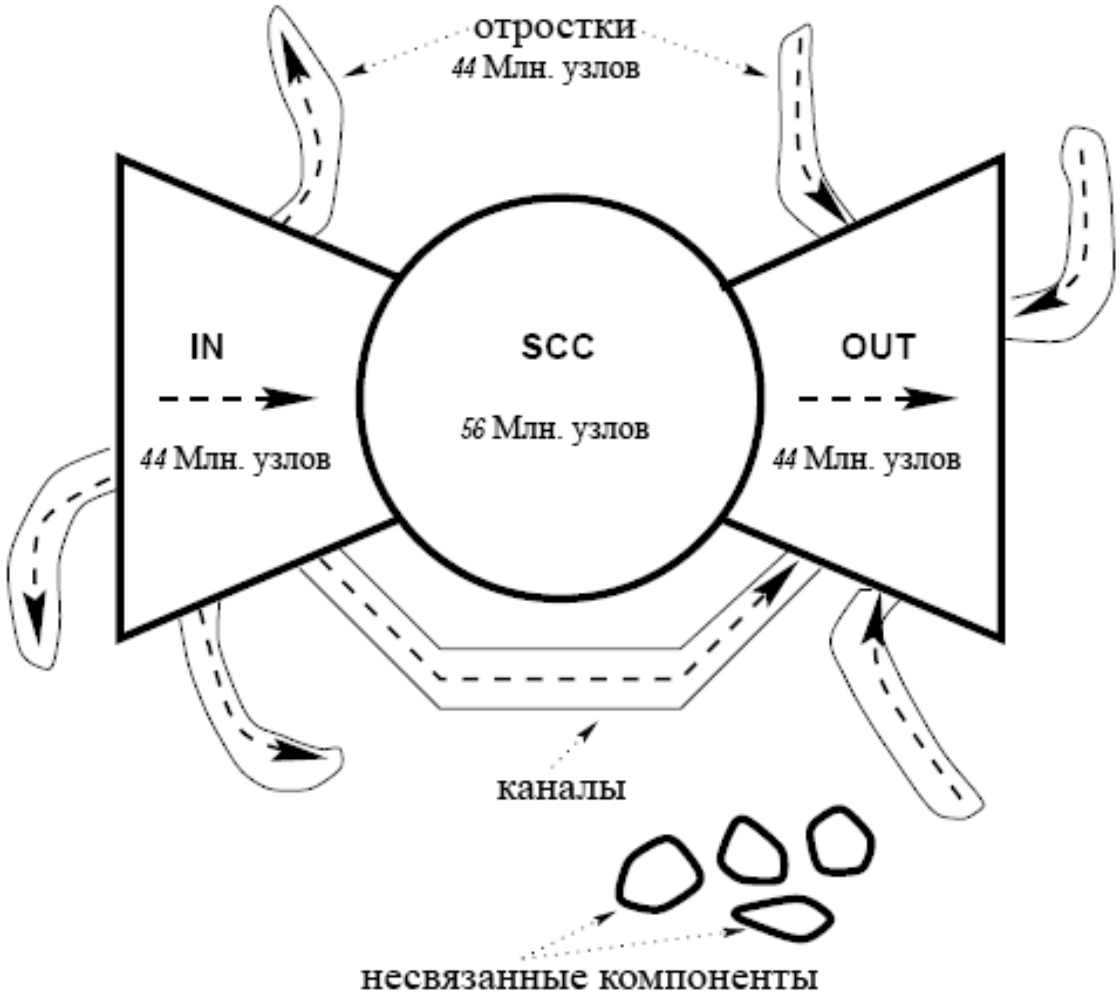
Модель Bollobas – Riordan, 2002:

– моделирование проведения m ребер из n-ой вершины.

Модель Barabashi – Albert, 2002:
– предпочтительное присоединение (“деньги к богатым”);
безмасштабные сети - лишь небольшое число вершин имеет большое число связей.

Модель Strogatz – Watts:
– сети с феноменом “тесного мира”.

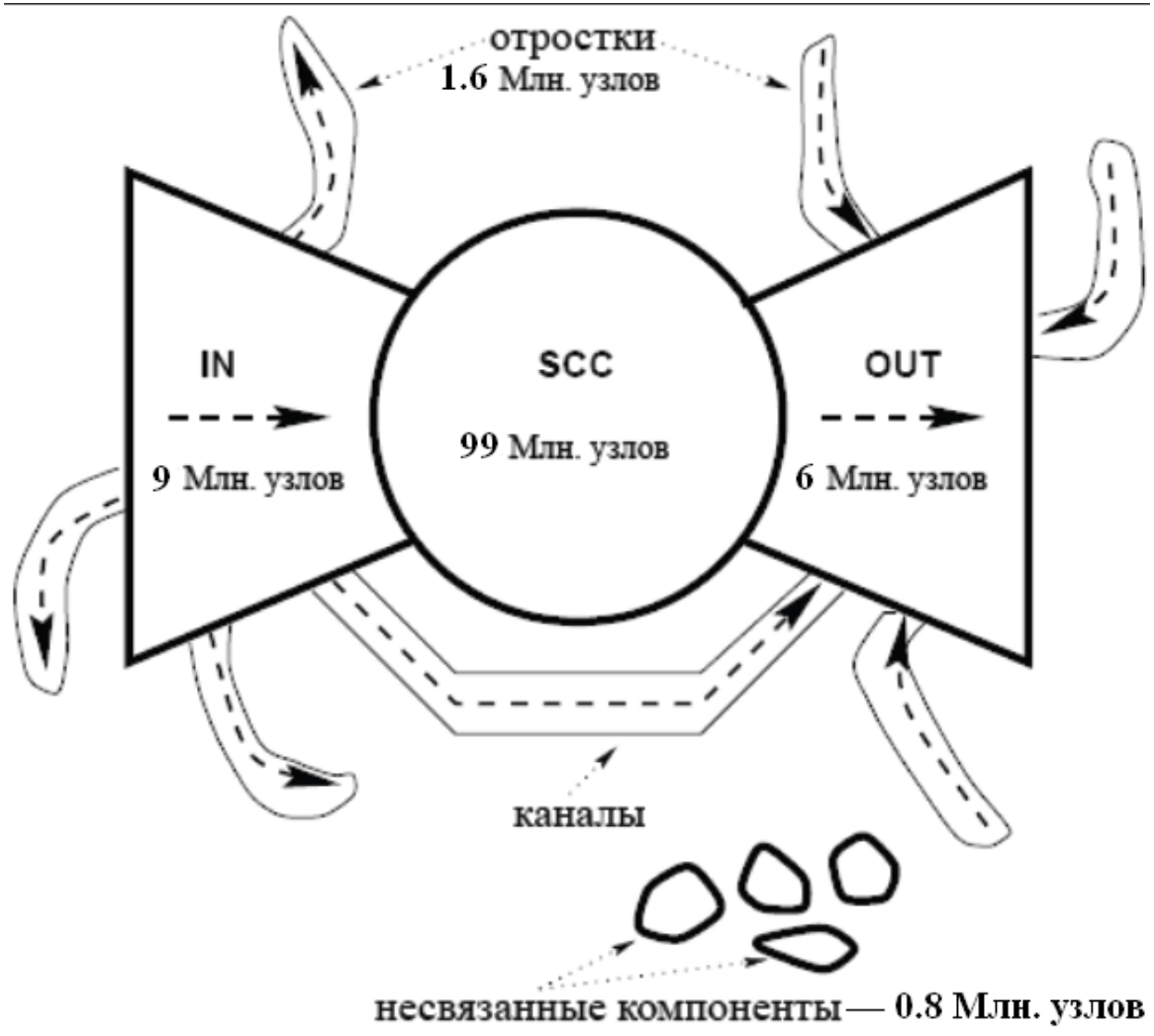
Модель “бабочка”, Broder, 1999:



SCC - сильно связная компонента

Han – Lee – Lee, 2007:

Анализ веб-графа корейской паутины в модели “бабочка”: вершин - 116 млн., связей - 2.7 млрд.



Сравнение размеров строго связной компоненты SSC (“ядра”):

Глобальная паутина	28%
Китайская паутина	80%
Корейская паутина	86%

Thelwell – Wilkinson, 2003:

Сравнительный анализ сети университетов в модели “бабочка”:

	Австралия	Новая Зеландия	Великобритания
вершин	3 511 507	305 180	6 533 600
связей	18 031 706	1 874 141	31 250 705
SCC	27%	30%	30%
OUT	73%	70%	70%

Имеющийся научный задел.

Имеющийся научный задел (ИВТ):
1998 г. Web Impact Factor (от Библиометрии к Вебометрии).

S – размер сайта (число страниц);
V – видимость (число внешних ссылок);
R – число “тяжелых” файло (pdf, doc, ppt);
Sc – индекс цитирования;

$$W = \alpha S + \beta V + \gamma R + \delta Sc.$$

Ю.И. Шокин, О.А. Клименко: Моделирование научной сети СО РАН и ее вебометрический анализ.
Регулярное обновление рейтинга сайтов всех институтов СО РАН:
http://www.ict.nsc.ru/ranking
$\alpha = 1$, $\beta = 2$, $\gamma = 2$, $\delta = 1,5$

Имеющийся научный задел (ИМ):

Количественные инварианты в исследованиях молекулярных графов и поиск общих частей графов (постулат "структура-свойство").

Молекулярные инварианты:
– физико–химические (молекулярный вес, мольный объем)
– квантово–химические (дипольный момент, энергия резонанса)
– геометрические (ван-дер-ваальсов объем)

Графовые инварианты:
– структурные (по наличию определенных фрагментов);
– топологические индексы (Рандича, Винера, Хосойя);
– информационно–теоретические (на основе формулы Шеннона с использованием дистанционных свойств графа и пр.).

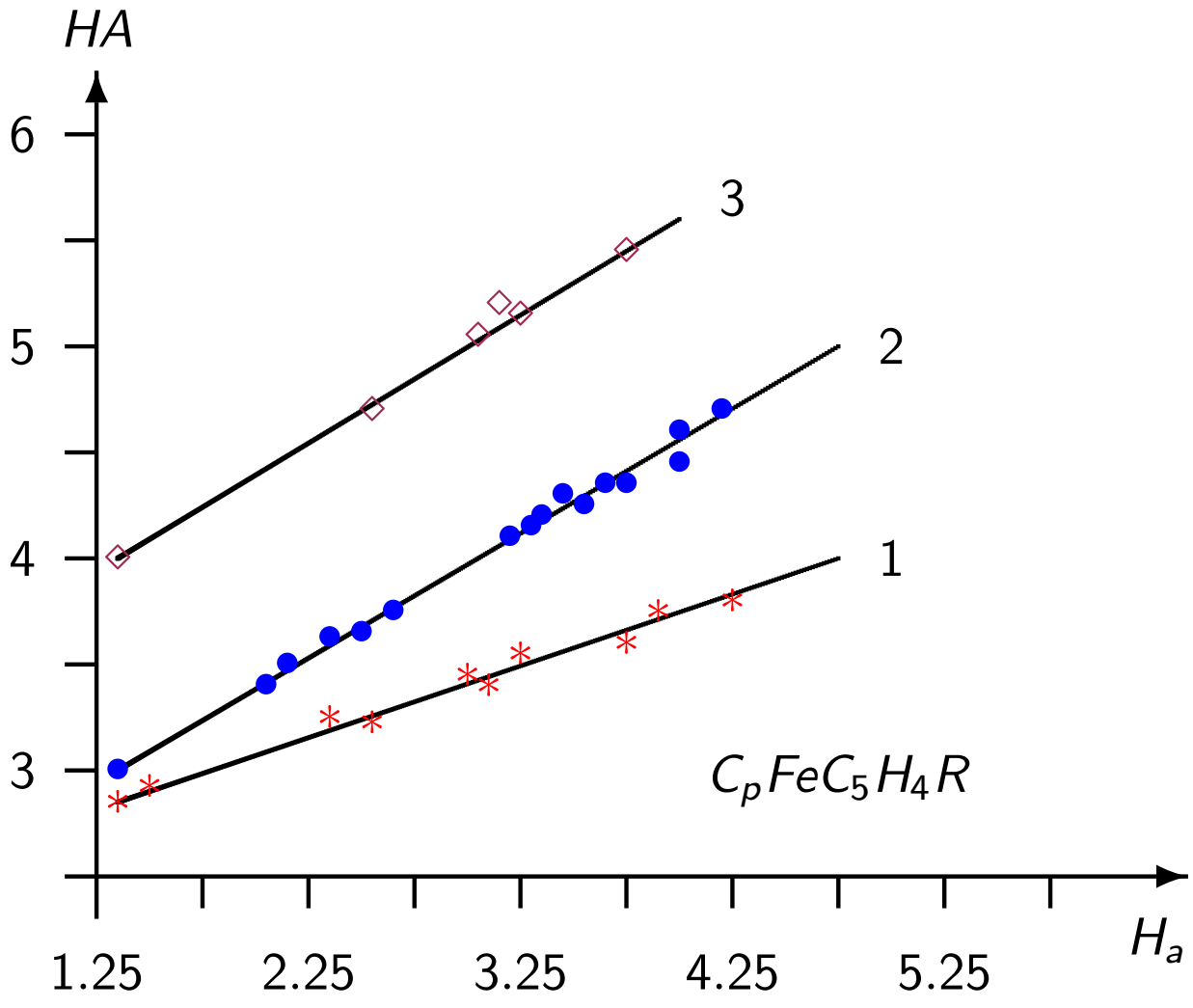
Информационно–теоретические инварианты обладают высокой дискриминирующей способностью. [Konstantinova, E.V.; Vidyuk, M.V. J. Chem. Inf. Comp. Sci. **43** 1860–1871 (2003)]

Пример. Производные ферроцена $C_pFeC_5H_4R$.

Найдены линейные корреляции между информационными индексами молекулярных графов (H_a) и информационными индексами масс–спектров (HA) соответствующих химических соединений:

$$1 \ : \ HA = 1.70 + 0.57\, H_a \quad (r = 0.940, n = 10)$$
$$2 \ : \ HA = 1.24 + 0.98\, H_a \quad (r = 0.975, n = 16)$$
$$3 \ : \ HA = 3.91 + 0.41\, H_a \quad (r = 0.940, n = 6)$$

Индекс корреляции: от 0.94 до 0.975.


HA
6
5
4
3
3
2
1
$C_pFeC_5H_4R$
1.25
2.25
3.25
4.25
5.25
H_a

Приложения:

в Институте элементоорганических соединений им. А.Н. Несмеянова РАН для поиска спектро–структурных корреляций и исследования связей структура–активность на примере (~ 20) органических и металлоорганических соединений.

В частности,

[Nekrasov, Yu.S.; Sukharev, Yu.N.; Tepfer, E.E.; Yakushin, S.: Electron impact mass spectra data processing for evaluation of gas–phase reactivity of cymantrene (tricarbonyl η5–cyclopentadienylmanganese) derivatives. Eur. J. Mass Spectrom. **8** 247–251 (2002)]

[Nekrasov, Yu.S.; Sukharev, Yu.N.; Tepfer, E.E.: Determination of spectrum–structure correlations based on integral parameters of mass–spectra. J. Analyt. Chem. **20** 1035–1037 (2005)]

Структурный анализ графов.

Поиск общих частей (общих подграфов) молекулярных графов. Реализовано для базы данных лекарственных соединений. Графы имеют метки на вершинах и веса на ребрах.

Таксономия лекарственных соединения по наборам характеристик, включающим структурные формулы.

План действий:

Теория:

a. Разработка и обоснование новых методов и моделей структурного и метрического анализа сложных сетей.

Практика:

b. Разработка программного обеспечения для известных и новых методов и моделей.

Приложения:

c. Структурный и метрический анализ веб-сети институтов СО РАН.

d. Структурный и метрический анализ биологических сетей. (Каких?)

e. Структурный и метрический анализ социальных сетей. (Каких?)