在线社交网络上用户社交行为的可预测性初探

许小可(xiaokeeie@gmail.com) 大连民族学院

2013/04/25

复杂网络研究的几个方面

- 分析
- 建模
- 预测
- 控制

预测很有用，无处不在

预测（**forecasting**）是预计未来事件的一门艺术，一门科学。它包含采集历史数据并用某种数学模型来外推与将来。它也可以是对未来的主观或直觉的预期。

预测的重要意义就在于它能够在自觉地认识客观规律的基础上，借助大量的信息资料和现代化的计算手段，比较准确地揭示出客观事物运行中的本质联系及发展趋势，预见到可能出现的种种情况，勾画出未来事物发展的基本轮廓，提出各种可以互相替代的发展方案，这样就使人们具有了战略眼光，使得决策有了充分的科学依据。

Reference:http://wiki.mbalib.com/wiki/%E9%A2%84%E6%B5%8B

预测可以赚钱：轮盘赌

AIP | Chaos
An Interdisciplinary Journal of Nonlinear Science

What your peers have been reading...

A Listing of the Most Read Articles in 2012 published in *Chaos*

Predicting the outcome of roulette
Michael Small and Chi Kong Tse
Chaos **22**, 033150 (2012)

预测可以救命：地震预测


四川雅安地震 余震不断
据中国国家地震台网测定
北京时间 4月20日
四 川
成都市
雅安
8: 06
震级 4.8级
震源 10公里
震中 北纬30.2 东经102.9
8: 07
震级 5.1级
震源 10公里
震中 北纬30.3 东经102.9
8: 31
震级 4.1级
震源 12公里
震中 北纬30.4 东经102.9
8: 37
震级 4.1级
震源 19公里
震中 北纬30.2 东经102.9
9: 02
震级 4.6级
震源 15公里
震中 北纬30.3 东经102.9
8: 02
震级 7.0级
震源 13公里
震中 北纬30.3 东经103.0
9: 39
震级 4.5级
震源 17公里
震中 北纬30.1 东经102.9
9: 37
震级 4.9级
震源 15公里
震中 北纬30.3 东经102.9
9: 26
震级 4.4级
震源 17公里
震中 北纬30.2 东经102.9
9: 20
震级 4.6级
震源 19公里
震中 北纬30.2 东经102.9
9:11
震级 4.3级
震源 16公里
震中 北纬30.2 东经102.8
截至9时40分
发生的4.0级以上余震
新华社记者 秦迎 编制

理想是。。。

实际上可能是。。。

问题：我们的预测靠谱吗？

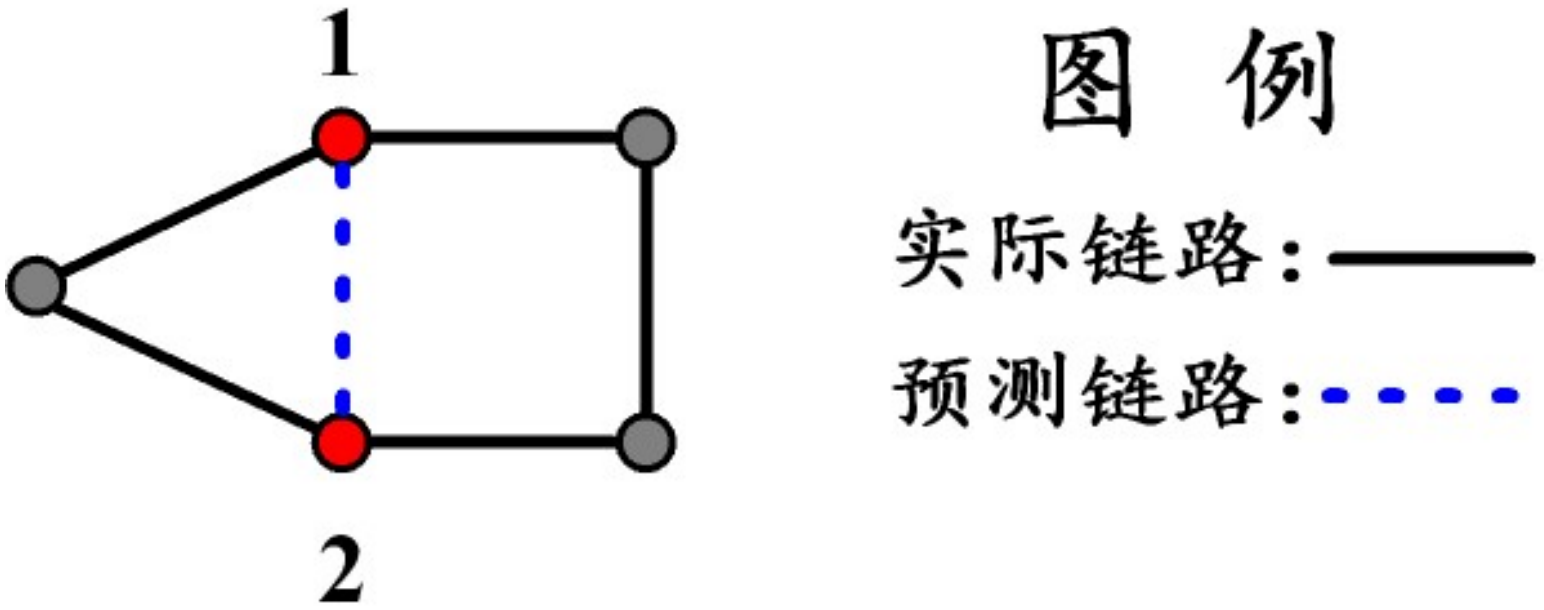
- 靠谱的程度：可预测性
- 能否得出某方法进行预测准确性的上限
- 能否从理论上比较各种预测方法的优劣
- 能否找出最具提高潜力的算法进行改进
- 能否找到最可预测的个体和群体加以利用

在线社交网络中的预测问题

- 结构网络上的链路预测和节点预测
- 功能网络上的行为预测和信息传播预测

1.结构网络中的链路预测

一种数据记录了用户在不同时间点上和其他用户建立好友关系或解除好友关系，这种数据描述了社交网络上用户之间显性关系（谁和谁是朋友）的演化，这种数据构成了社交网络上用户的关系网络，我们称之为社交结构网络。



常见的局域链路预测算法

表1　10种基于节点局部信息的相似性指标

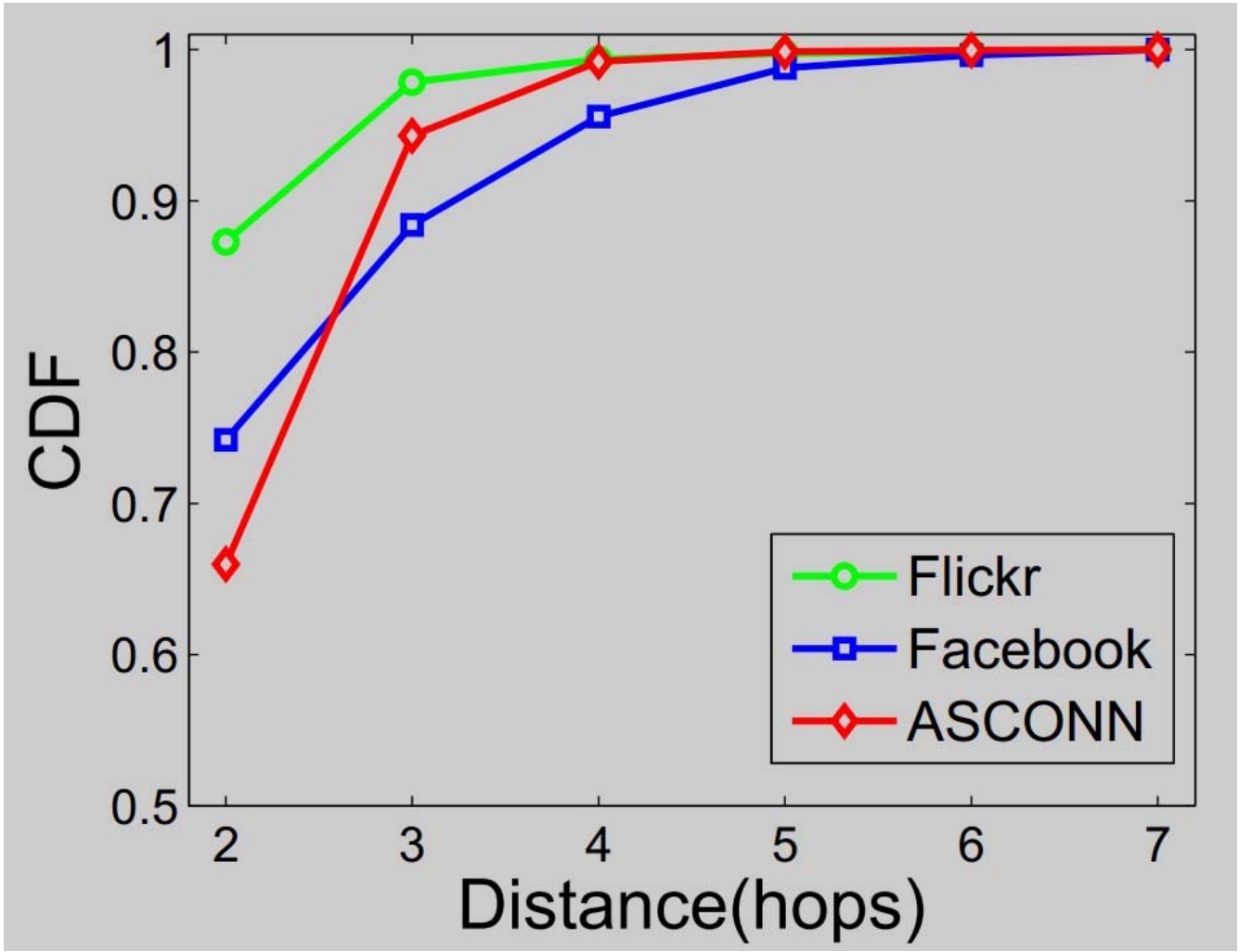
名称	定义	名称	定义
共同邻居(CN)	$s_{xy}=\lvert\Gamma(x)\cap\Gamma(y)\rvert$	大度节点不利指标(HDI)	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\max\{k(x),k(y)\}}$
Salton指标[27]	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\sqrt{k(x)\times k(y)}}$	LHN-I指标[22]	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{k(x)\times k(y)}$
Jaccard指标[28]	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$	优先链接指标(PA)[31]	$s_{xy}=k(x)\times k(y)$
Sorenson指标[29]	$s_{xy}=\frac{2\lvert\Gamma(x)\cap\Gamma(y)\rvert}{k(x)+k(y)}$	Adamic-Adar指标(AA)[32]	$s_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\lg k(z)}$
大度节点有利指标(HPI)[30]	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\min\{k(x),k(y)\}}$	资源分配指标(RA)[33]	$s_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{k(z)}$

可比较各种链路预测算法的优劣

表3　10种基于节点局部信息的相似性在6个网络链路预测中的精度比较

Index	PPI	NS	Grid	PB	INT	USAir
CN	0.889	0.933	0.590	0.925	0.559	0.937
Salton	0.869	0.911	0.585	0.874	0.552	0.898
Jaccard	0.888	0.933	0.590	0.882	0.559	0.901
Sorensen	0.888	0.933	0.290	0.881	0.559	0.902
HPI	0.868	0.911	0.585	0.852	0.552	0.857
HDI	0.888	0.933	0.590	0.877	0.559	0.895
LHN-I	0.866	0.911	0.585	0.772	0.552	0.758
PA	0.828	0.623	0.446	0.907	0.464	0.886
AA	0.888	0.932	0.590	0.922	0.559	0.925
RA	0.890	0.933	0.590	0.931	0.559	0.955

能否可出这些预测方法的上限？



如何改进算法

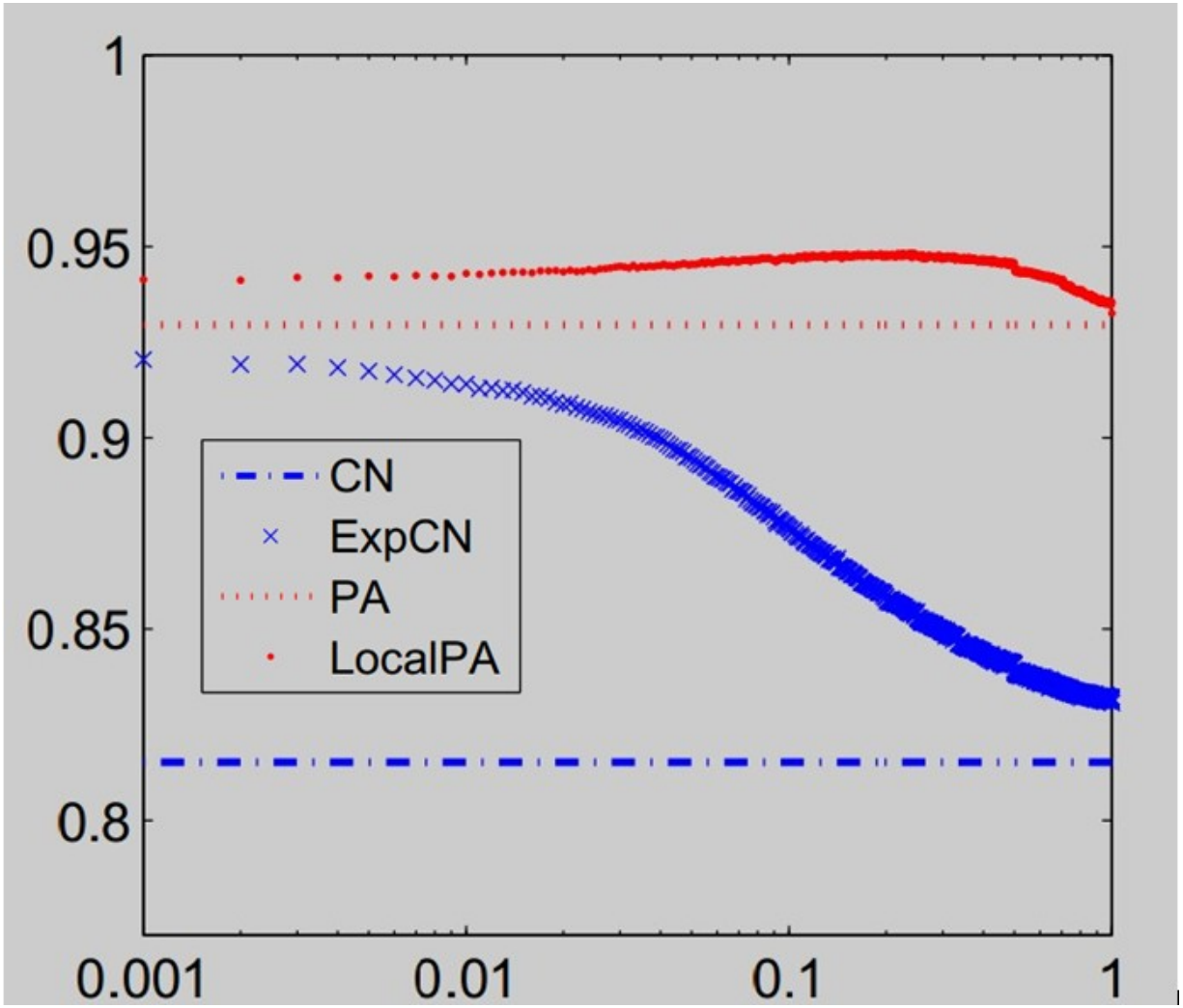
ExpCN (Expanded Common Neighbor) and *LocalPA* (Local Preferential Attachment):

$$S^{ExpCN} = A^2 + \alpha A^3 + \alpha^2 A^4, \tag{7}$$

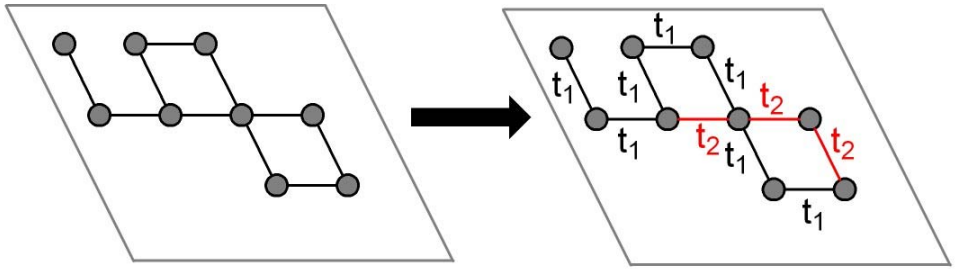
and

$$s_{xy}^{LocalPA} = |\Gamma(x)| \times |\Gamma(y)| \cdot \alpha^{f(x,y)}. \tag{8}$$

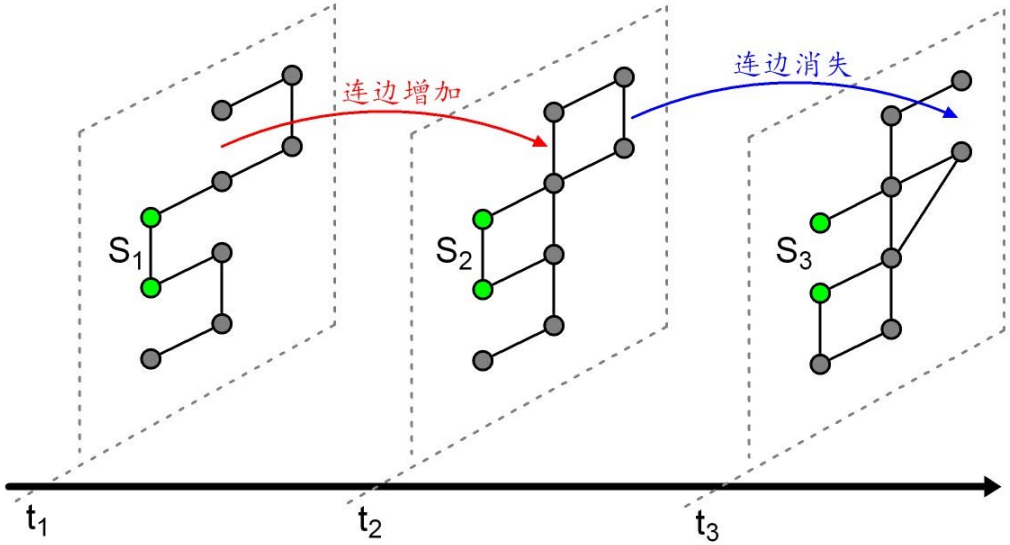
改进后链路预测的结果



进一步的改进：加入时间信息



(a) 无权结构网络　　(b) 带有链接时间序列的结构网络

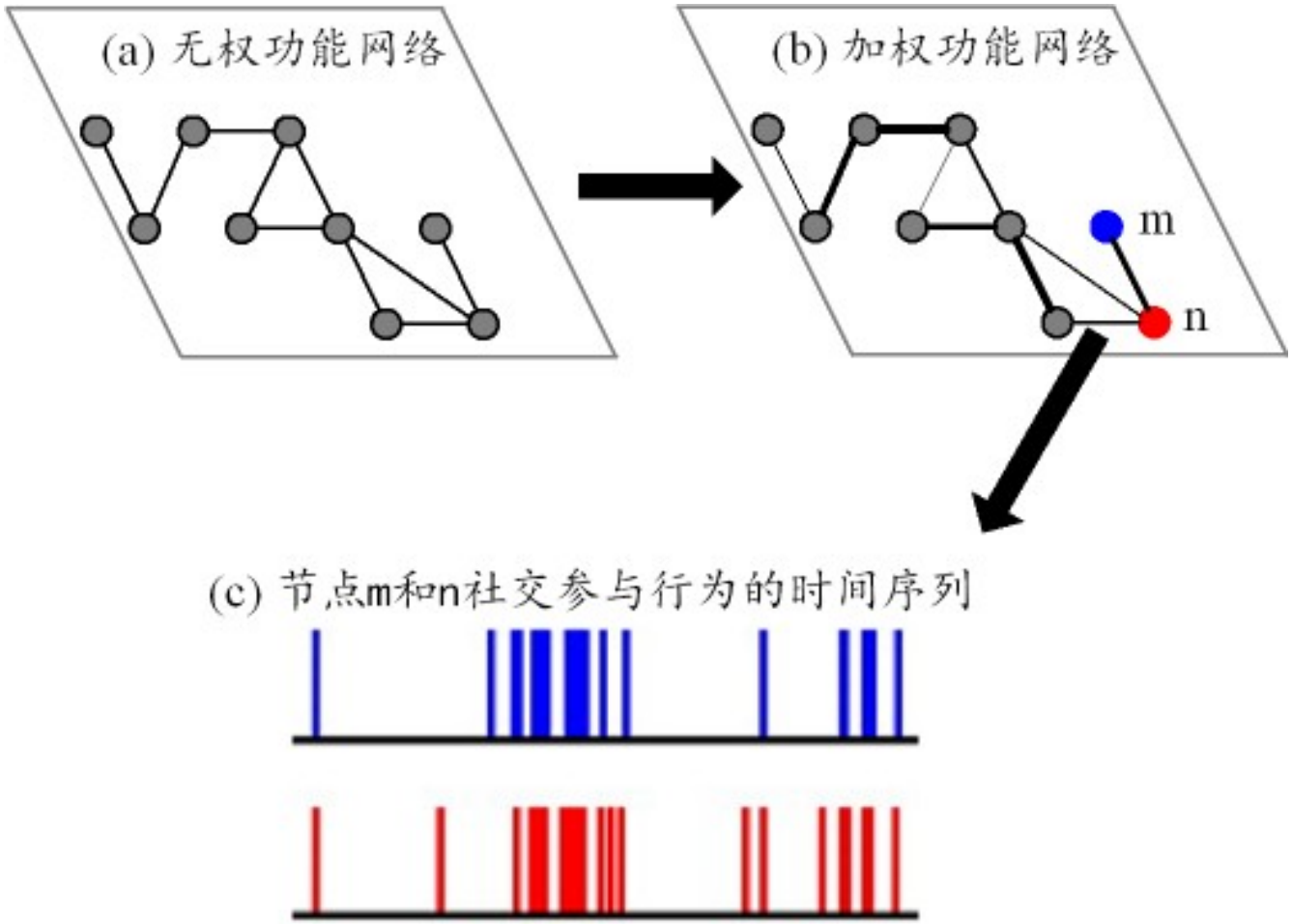


(c) 多时间片结构网络

2.功能社交网络的行为预测

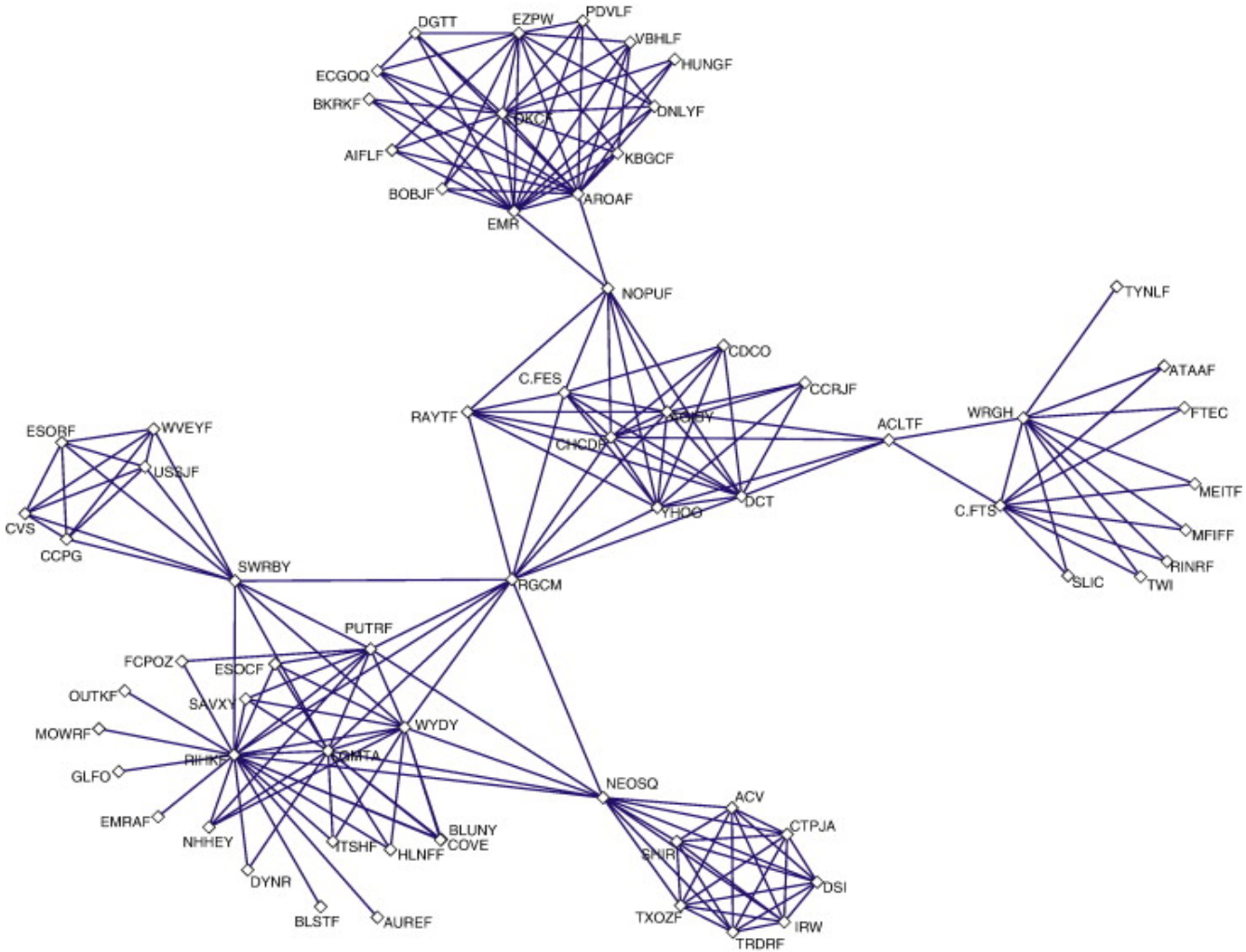
一种数据记录了用户在不同时间点上与其他用户共同参与某项调查、发表微博、收发短消息、语音聊天、分享文件等行为，这种数据描述了社交网络上用户各种参与行为的演化，完成了用户的社交功能，这种数据构成了社交功能网络。

功能社交网络

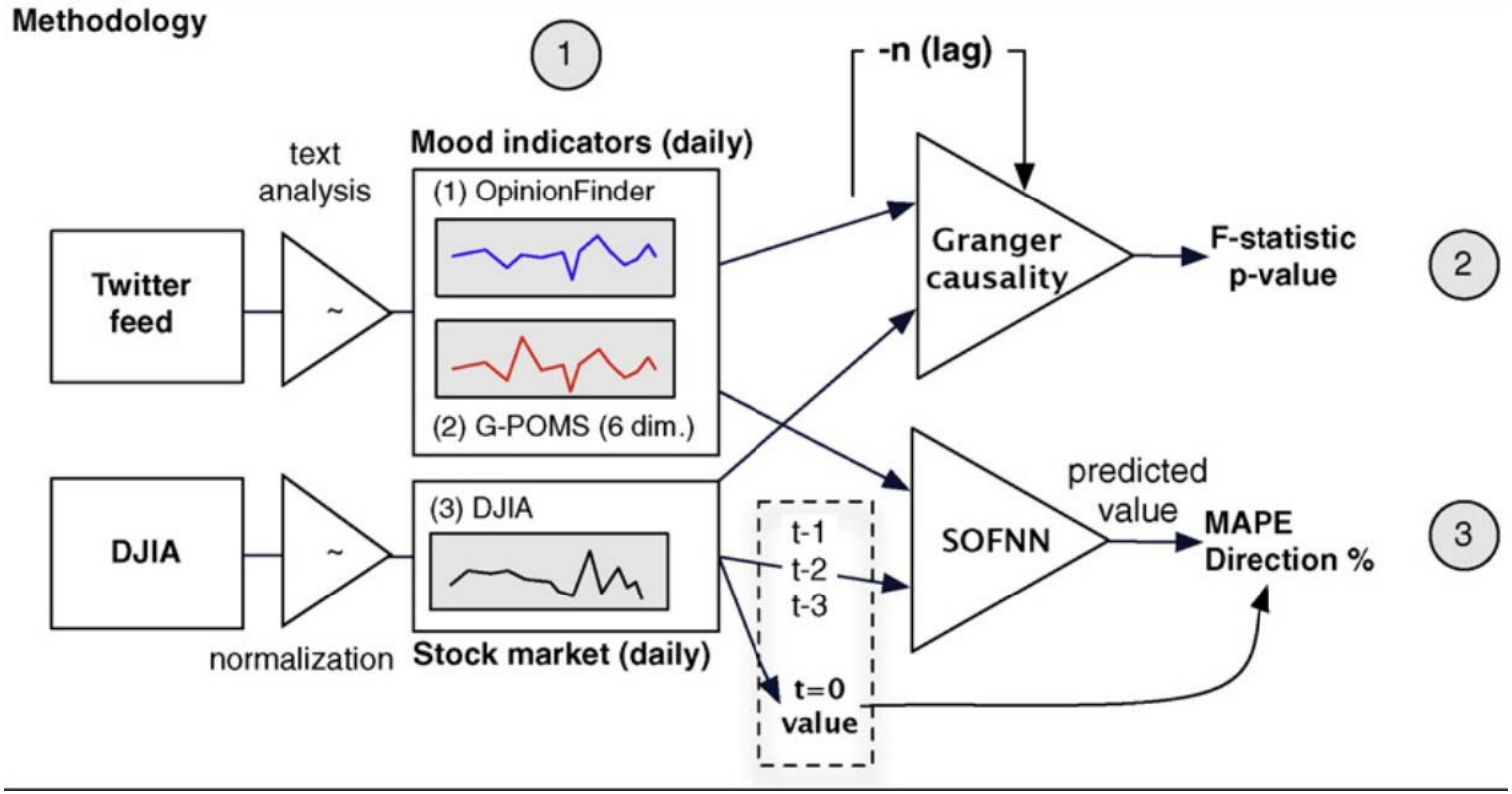


股票数据：单变量时间序列预测

基于复杂网络的时间序列预测



基于多个网络的时间序列预测



Reference: Journal of computational science, 2:1-8, 2011.

符号时间序列中的预测问题

- 能否预测出用户下一次社交行为的对象
- 能否根据行为时间序列重构出社交关系

预测下一次社交行为的对象

人类移动模式的**93%**是可以预测的

基于信息熵理论的方法

1. 随机熵

$$H_i^0 \equiv \log_2 k_i,$$

2. 非相关熵

$$H_i^1 \equiv -\sum_{j \in \mathcal{N}_i} P_i(j) \log_2 P_i(j),$$

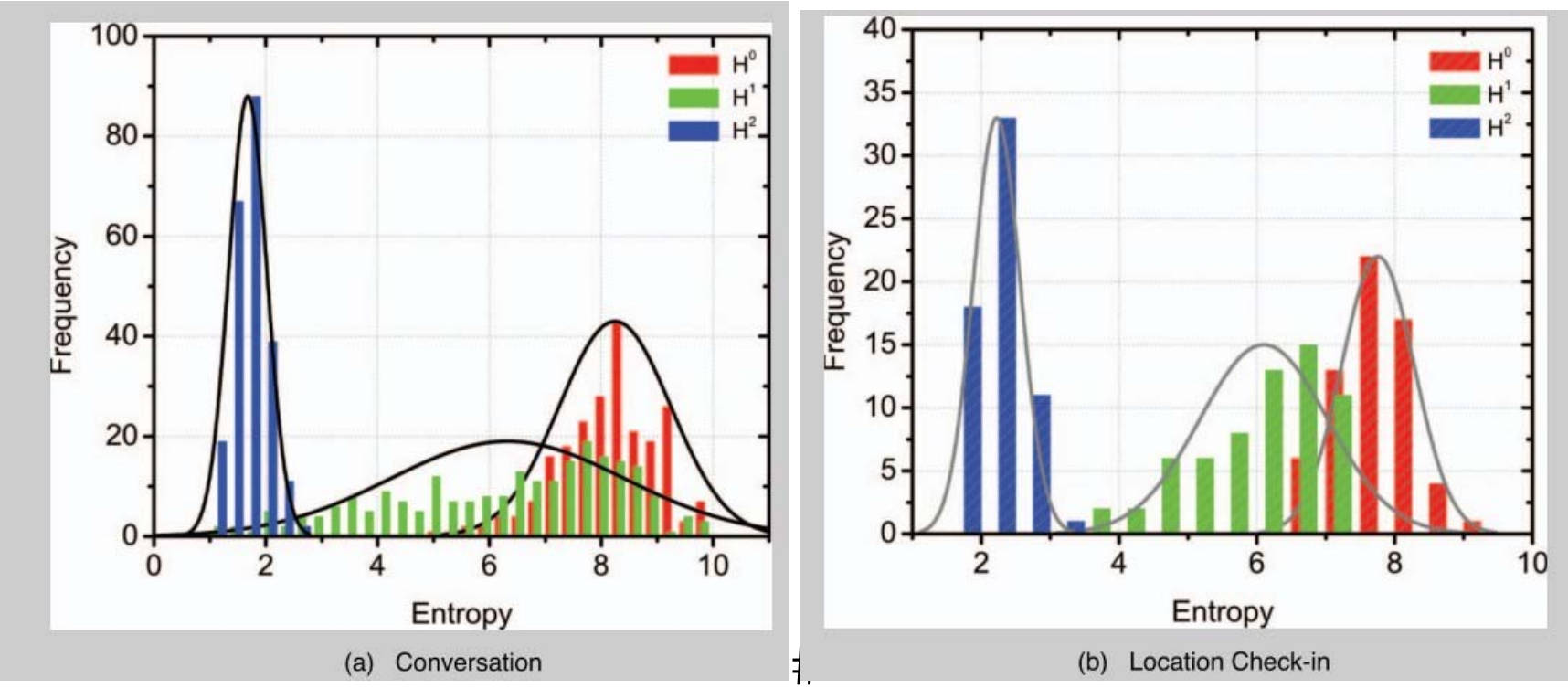
3. 条件熵

$$H_i^2 \equiv -\sum_{j \in \mathcal{N}_i} P_i(j) \sum_{\ell \in \mathcal{N}_i} P_i(\ell|j) \log_2 P_i(\ell|j),$$

4. 互信息

$$I_i \equiv H_i^1 - H_i^2 = \sum_{j,\ell \in \mathcal{N}_i} P_i(\ell, j) \log_2 \frac{P_i(\ell, j)}{P_i(\ell) P_i(j)},$$

基于信息熵理论的可预测性



Reference: Sci.Rep. 2:633, 2012.

基于转移熵理论重构社交关系

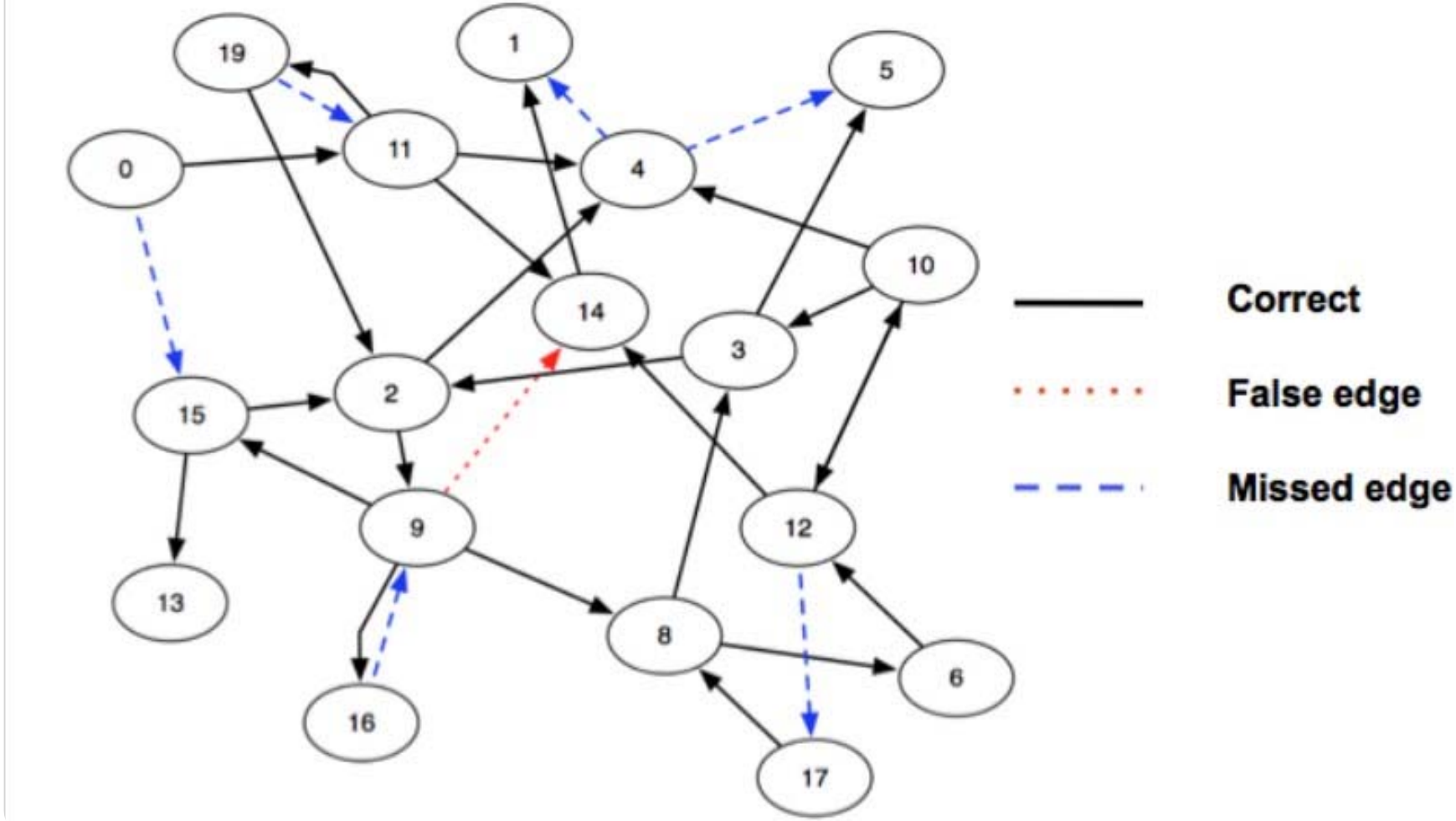
Information transfer can also be written as the mutual information between Y's present and X's past, conditioned on Y's past.

$$T_{X \to Y} = H(Y_t : X_{t-1}^{(t-k)} | Y_{t-1}^{(t-k)})$$

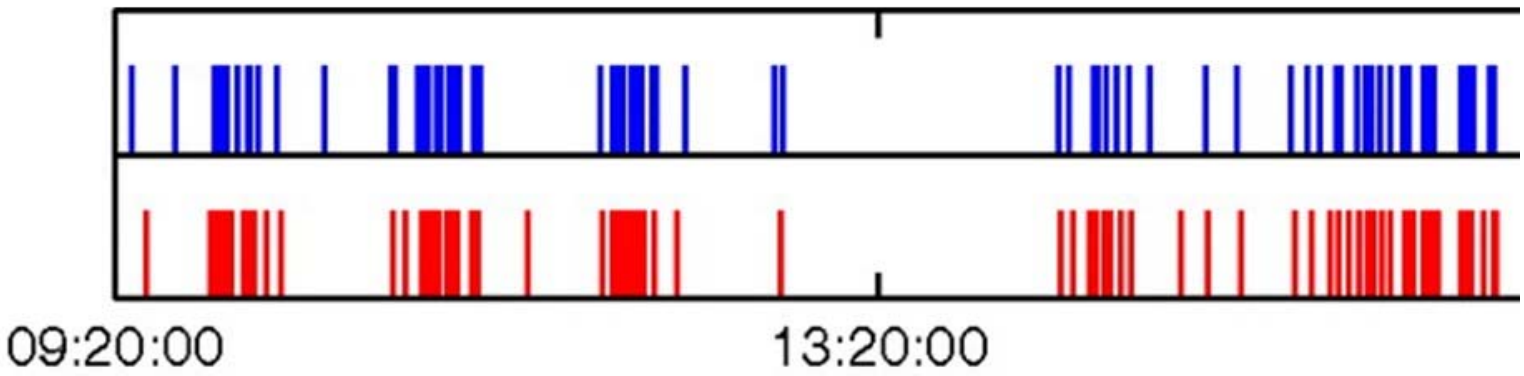
Because of the conditioning on Y's past, the transfer entropy is asymmetric, as opposed to standard mutual information, and thus better suited for characterizing directed information transfer. This captures the intuition that we are only interested in information about Y that is explained by X but cannot be explained by Y's own history.

Reference: WWW 2012 Session: Information Diffusion in Social

基于转移熵理论重构社交关系



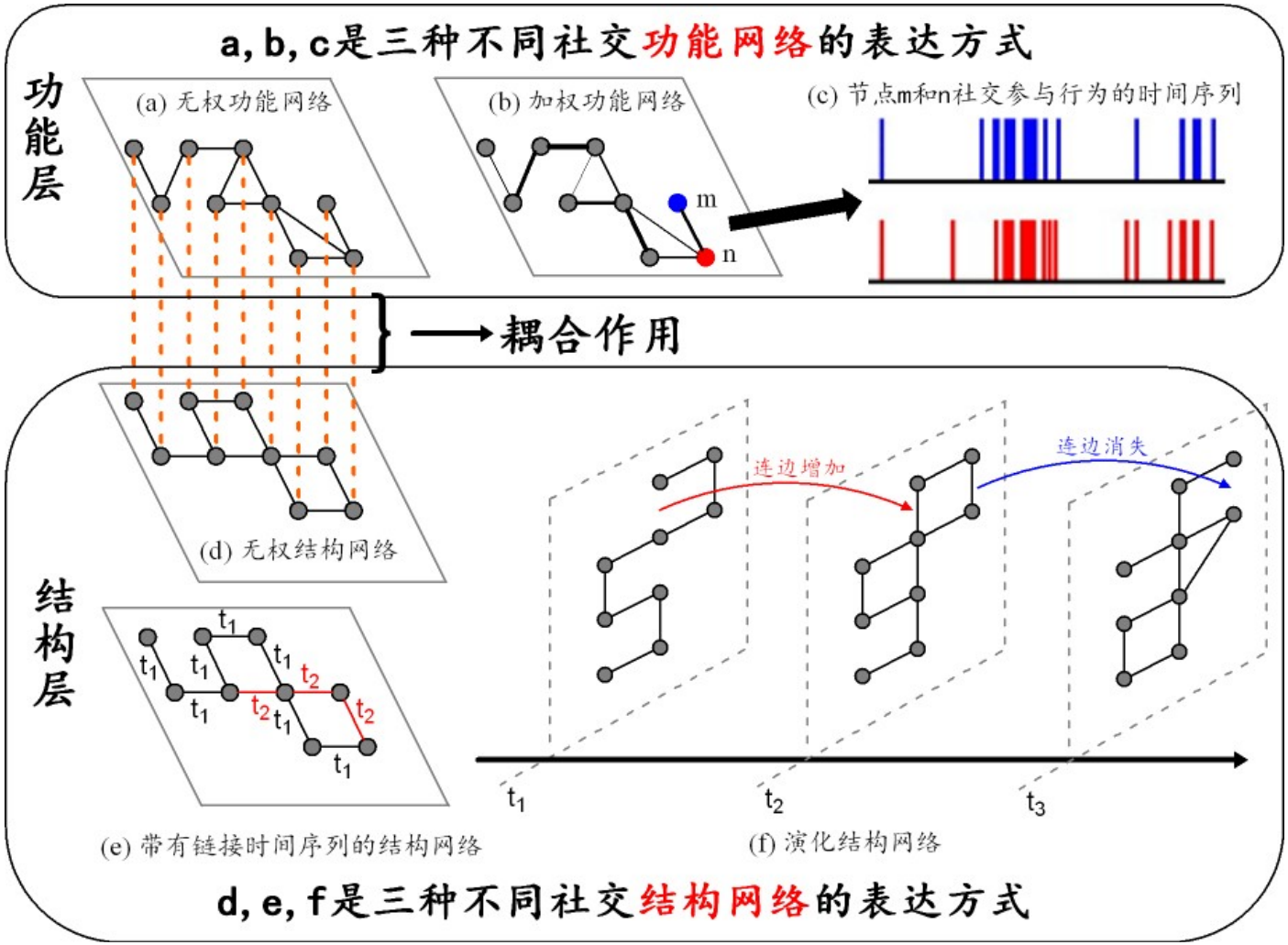
基于行为同步重构社交关系



可准确识别出活跃个体的强链接关系，非活跃用户的社交关系无法识别。个体的差异性导致行为可预测性的差别，鉴于此能否

- 将对整体预测的研究改为对某类可预测性强群体的预测
- 根据个体可预测性的差异来研究整体中特定群体的预测

双层耦合社交网络：结构-功能



问题和讨论