网络科学

第一讲

2019版

罗铁坚

课程：https://tjluo-ucas.github.io/ns/

中国科学院大学

提纲

- 科研范式
- 内容提要
- 基本概念
- 网络结构
- 机器学习与网络科学
- 总结

经典科学研究范式

	数学 Math	科学 Science	工程 Engineering
1 问题提出	刻画要研究的对象，给出合理的（定义）	观察可能重复的现象或模式，做出（假设）	提出期望构建系统的行为和响应（需求）
2 形成概念	假设研究对象之间存在某种关系（定理）	构造模型来解释观察到的现象并使其能进行预测（模型）	建立形式描述要建系统的功能和交互行为（规格说明）
3 解决问题	证明	收集数据 实验验证	设计与实现原型
4 验证结果	解释结果	解释结果	测试原型
5 实际应用	应用	预测将要发生事件	建造

信息科学研究范式

Computing **计算** （提升智力）

1 问题提出	观察和思考现实系统中的信息处理过程的表达。
2 形成概念	发现计算模型（算法和数据）产生的系统行为。
3 解决问题	算法转化为计算机程序，来实现设计要求。
4 验证结果	验证计算机程序运行是否正确、安全、可靠、高性能。
5 实际应用	运行结果是否反映实际需求，持续监视和评估系统。

若关注理论或数学，则有形式定义和定理的讨论；
若偏重发现规律，则有相关的假设、模型和实验验证；
若设计产品或提供服务，则有规范说明和工程实施内容。

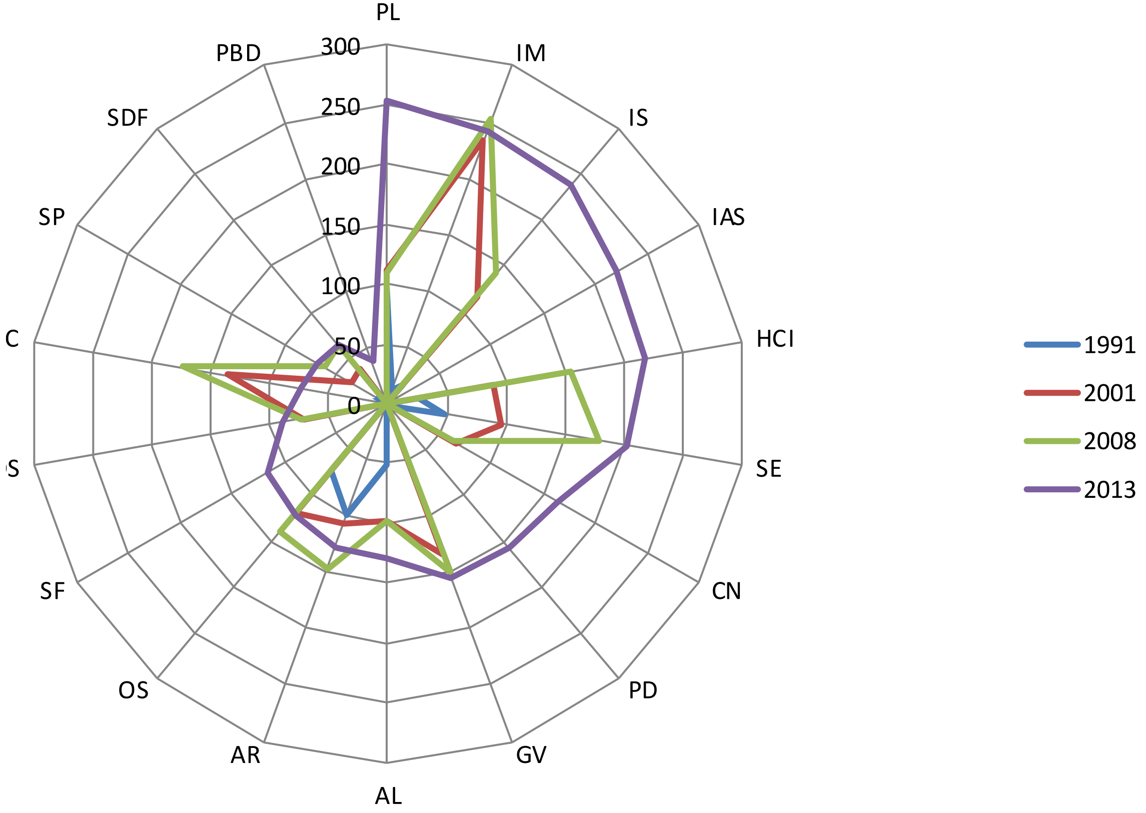
Methodology of Modeling and Analyzing Disciplinary Evolution –

《工程研究--跨学科视野中的工程》

2013年18个知识域列表

- AL - Algorithms and Complexity
- AR - Architecture and Organization
- CN - Computational Science
- DS - Discrete Structures
- GV - Graphics and Visualization
- HCI - Human-Computer Interaction
- IAS - Information Assurance and Security
- IM - Information Management
- IS - Intelligent Systems
- NC - Networking and Communications
- OS - Operating Systems
- PBD - Platform-based Development
- PD - Parallel and Distributed Computing
- PL - Programming Languages
- SDF - Software Development Fundamentals
- SE - Software Engineering
- SF - Systems Fundamentals
- SP - Social Issues and Professional Practice

摘要：学科概念知识图谱是判断学科发展演化的重要依据,也是构建智能导学系统的基础性工作。本文提出了一种基于概念的高精度层级结构模型,该模型通过抽取学科的知识体系中的基本概念,建立了一个学科知识体系数据库,使用自然语言处理和信息可视化等工具,描绘出该学科的演化发展脉络。为了验证模型和方法的有效性,我们选用了50多年来ACM/IEEE发布的计算机学科知识体系作为数据集进行分析,给出了6类学科知识图谱,这些图谱客观反映了计算机学科经历了知识领域体系重构和快速更新的发展历程。



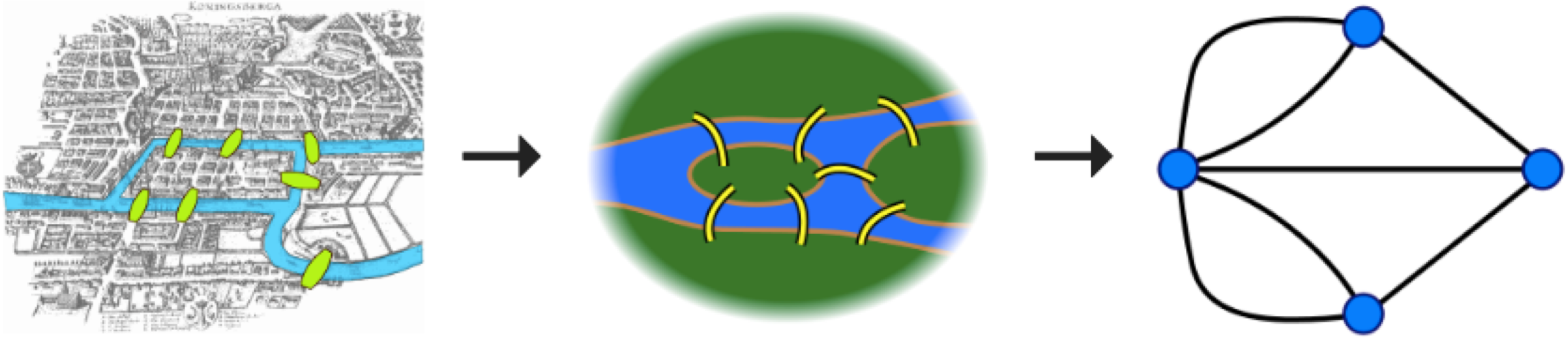
图：知识域5次调整变化的雷达图。60年有近10次教学大纲调整，图中选了近年的5次作为观察分析

网络科学

面向社会、技术或生物等复杂系统建模和分析的工具和方法。

- 从宏观宇宙到微观粒子的自然界、从人类全体到具体个人的社会都表现出“网络”的形态。数学、物理学、生物学、计算机科学、社会科学等学科专家都分别从各自关注点和角度对其进行研究，取得了有影响的研究成果。他们对网络中的个体或整体的形成、演化，节点间的相互影响等共性课题表现了极大兴趣。
- 我们需要建立一门对复杂现象进行网络建模和计算推理来探索事物运作机理并预测其发生和发展趋势的学科。

学科简史



- 1736年欧拉解决著名七桥不重复游走问题（论文 “The solution of a problem relating to the geometry of position”）创立了图论。
- 1878年 Sylvester 把网络应用到化学领域。
- 1933年 Jacob Moreno 把图论应用到心理学，发展了社会网络分析方法。
- 1936年 Dénes Kőnig 写出了第一本图论的论著。
- 1959年 Paul Erdős和Alfréd Rényi 把概率引入图论，提出了随机图网络理论。
- 1998年 David Krackhardt和Kathleen Carley 提出了元网络PCANS模型分析组织行为。
- 1998年 Watts等 提出小世界网络理论。”Collective dynamics of ‘small-world’ networks” . Nature. 393 (6684): 440–442。
- 1999年 Barabási等 提出无标度网络。”Emergence of scaling in random networks”. Science. 286 (5439): 509–512
- ……
- 虽然网络科学取得了丰富多彩的知识成果，但还有许多理论和应用问题等待我们挑战。

复杂系统是普遍存在的事物和现象

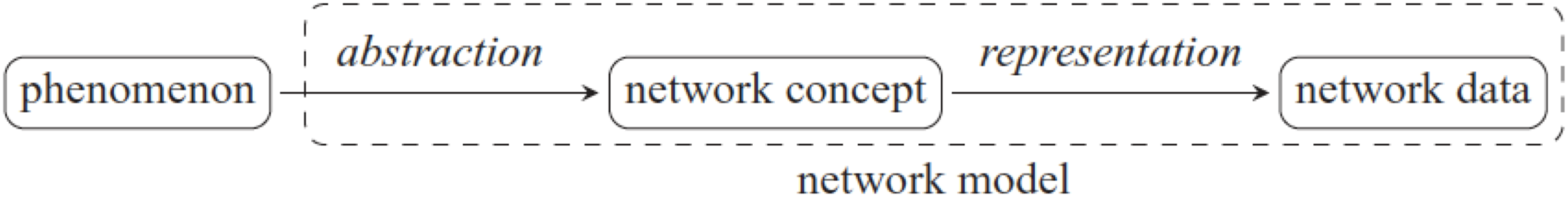
- 人类共有近70亿个人。
- 全球通讯系统把电子设备联系起来。
- 用链接把信息和知识组织起来形成网络。
- 基因和蛋白质相互作用模式构成了生命。
- 大脑中几十亿的神经元构成的网络隐藏着丰富的思想。

我们如何表达这些复杂系统呢?

中国科学院大学

内容提要

- “处处皆网络” 网络是人们普遍使用的概念。更精确地说，图是一个包含若干节点和联接节点的（有向或无向）边所构成的集合。网络则是节点或边被赋值后的图所构成的集合。人们观察具体现象，把研究对象抽象表示成网络模型，然后对网络模型中的元素赋予相应的数值（通过实际观察和度量）；使用相应算法对网络数据进行推理演算，得出结果，回答或解析具体现象（研究对象）的问题。



研究方法

基本假设：

1、社会、自然或技术“系统”是由不同部分有机组成。

2、这些关联的部分是存在某种“结构”的。

3、个体部分之间通过“结构”链接进行交互使系统得以运行。

4、系统内部的个体之间相互依存，并且任何个体的行为结果潜在地依赖其他个体的联合行为。

5、利用图论的知识来探讨系统的网络结构；研究个体节点的行为规律，采用博弈论的语言来建立基本模型。

研究方法：

1、确定研究问题（事件发生的原因或预测可能发展的趋势）

2、建模：构建网络图（节点和边的定义，需要领域知识）

3、分析：节点的行为模式或变化，全图的演化。

4、归纳：模式总类、发生的原因、调控手段

5、实验仿真和实际应用：收集什么数据、进行什么分析、给出何种决策。

讨论主题

1. 如何对网络结构进行分类；
2. 提出了什么形式化工具和方法刻画和度量网络结构；
3. 针对专业领域的网络数据，研发了哪些相关的算法和软件工具；
4. 研究了什么数学模型或机器学习模型来预测网络行为，理解网络形成和演化趋势；
5. 应用案例，如中国历代人物网络分析、社交网络流行机理、搜索广告和计算机网络等。

授课对象

计算机科学和软件工程专业领域的硕士或博士研究生，也欢迎其他学科领域关心数据科学的研究生。

预备知识

中文、英文、Python 程序设计语言；建议选课学生通过网络开放课程来温习这些知识。

评分标准

（1）2次作业 (50%)

（2）个人表现（团队协作、课堂提问或回答问题、点评小组课题等)(10%)

（3）2人小组课题 (40%)

候选课题

（1）网络搜索（Web Search： Ch. 14.4-14.5）

（2）搜索的付费广告市场（Sponsored Search as a Market： Ch 15.1-15.5）

（3）信息级联（Information Cascades： Ch. 16.1-16.7）

（4）网络效应，级联行为（Network Effects, Cascading Behavior： ch. 17.1-17.3, 19.1-19.4）

（5）富者更富现象（Rich-get-richer： ch 18.1-18.6）

（6）小世界理论（Small Worlds： Ch. 20.1-20.6）

（7）流行发生与演化（Epidemics： Ch 21.1-21.4, 21.6）

（8）若在这个选题之外，提供相关文献，提前打印或分发给大家

讲课和答疑安排

课程共20个学时，连续10周的课堂讨论和自主学习。

上课时间: 周四 8:30am - 10:00am ;上课地点: 国科大雁栖湖校区教1-113

答疑时间: 周四 3:00pm - 4:00pm；答疑地点: 国科大雁栖湖校区学园2-485；负责答疑：罗铁坚（教授），邮件:tjluo@ucas.ac.cn，

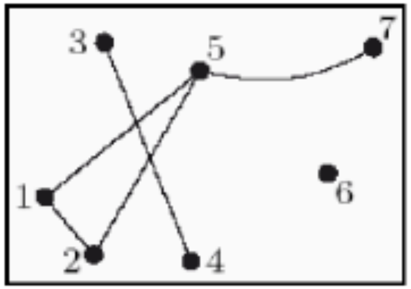
答疑时间: 周四 2:00pm - 5:00pm；答疑地点: 国科大雁栖湖校区学园2-505；负责答疑：刘 丹（助教），邮件:diananini@163.com

概念和术语（1）

Network

- A collection of individual or atomic entities
- Referred to as nodes or vertices (the “dots” or “points”)
- Collection of links or edges between vertices (the “lines”)
- Links can represent any pairwise relationship
- Links can be directed or undirected
- Network: entire collection of nodes and links
- For us, a network is an abstract object (list of pairs) and is separate from its visual layout
- that is, we will be interested in properties that are invariant
  - structural properties
  - statistical properties of families of networks

What different kinds of networks exist in the rea world?



概念和术语（2）

- Network size: total number of vertices (denoted N)
- Maximum number of edges (undirected): N(N-1)/2 ~ N^2/2
- Distance or geodesic path L between vertices u and v:
  - number of edges on the **shortest path** from u to v
  - can consider directed or undirected cases
  - infinite if there is no path from u to v
- Diameter of a network
  - worst-case diameter: largest distance between a pair
  - Diameter: longest shortest path between any two pairs
  - average-case diameter: average distance
- If the distance between all pairs is finite, we say the network is connected; else it has multiple components
- Degree of vertex v: number of edges connected to v
- Density: ratio of edges to vertices

图论与网络理论区别

- Graph Theory
  - Mathematics of graphs
  - Networks with pure structure with properties that are fixed over time
  - Focus on syntax rather than semantics
    - Nodes and edges do not have semantics
    - E.g. A node does not have a social identity
  - Concerned with characteristics of graphs
  - Proofs
  - Algorithms

Network Theory

- Relate to real-world phenomena
  - Social networks
  - Economic networks
  - Energy networks
- Networks are *doing something*
  - *Making new relations*
  - *Making money*
  - *Producing power*
- Are dynamic
  - Structure: Dynamics of the network
  - Agency: Dynamics in the network
- *Are active, which effects*
  - *Individual behavior*
  - *Behavior of the network as a whole*

社会网络

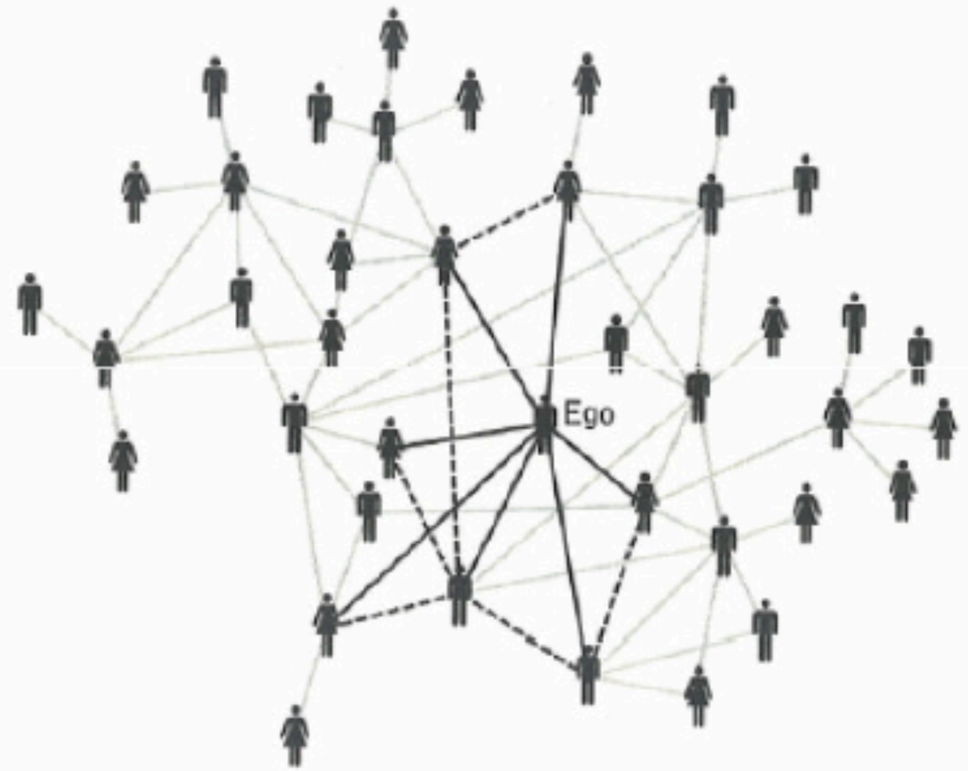
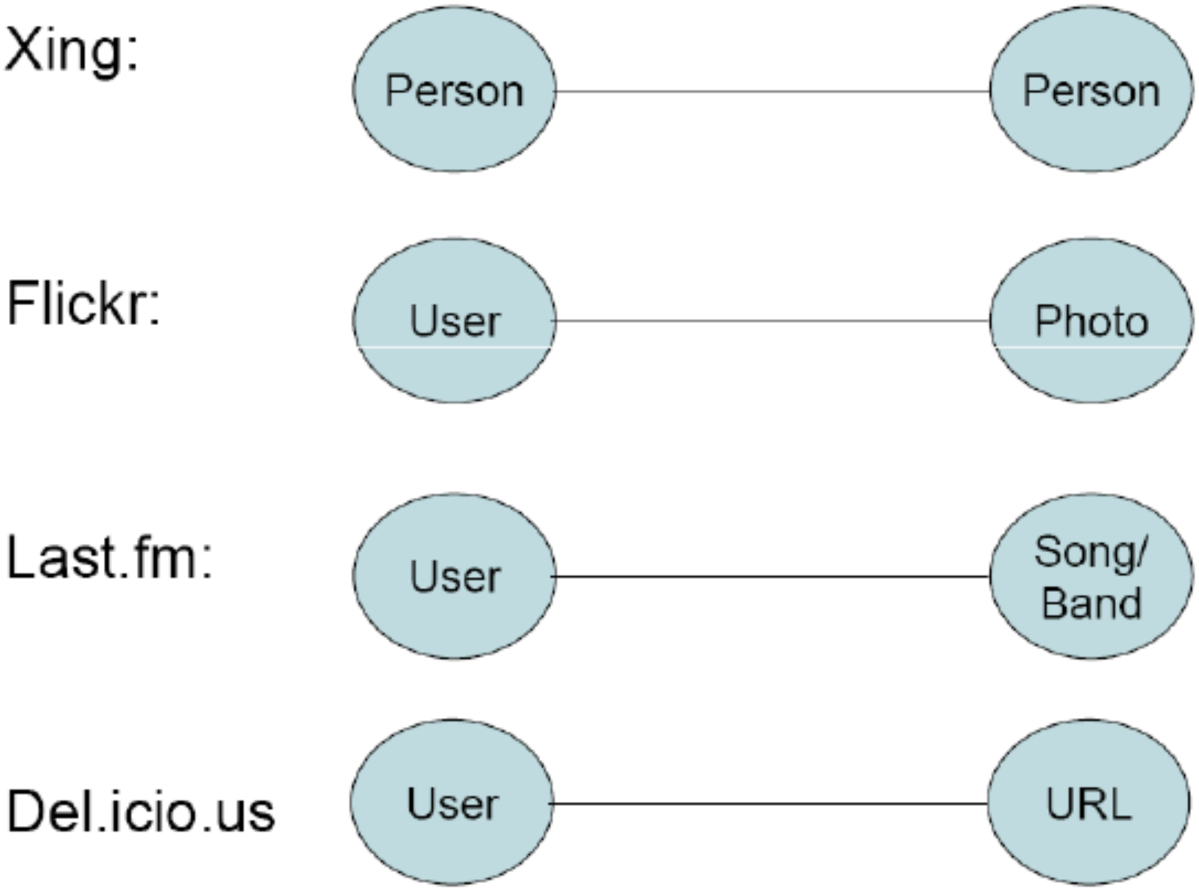


Figure 1.3. Real social networks exhibit clustering, the tendency of two individuals who share a mutual friend to be friends themselves. Here, Ego has

社会网络建模例子



Xing:

Person

Person

Flickr:

User

Photo

Last.fm:

User

Song/ Band

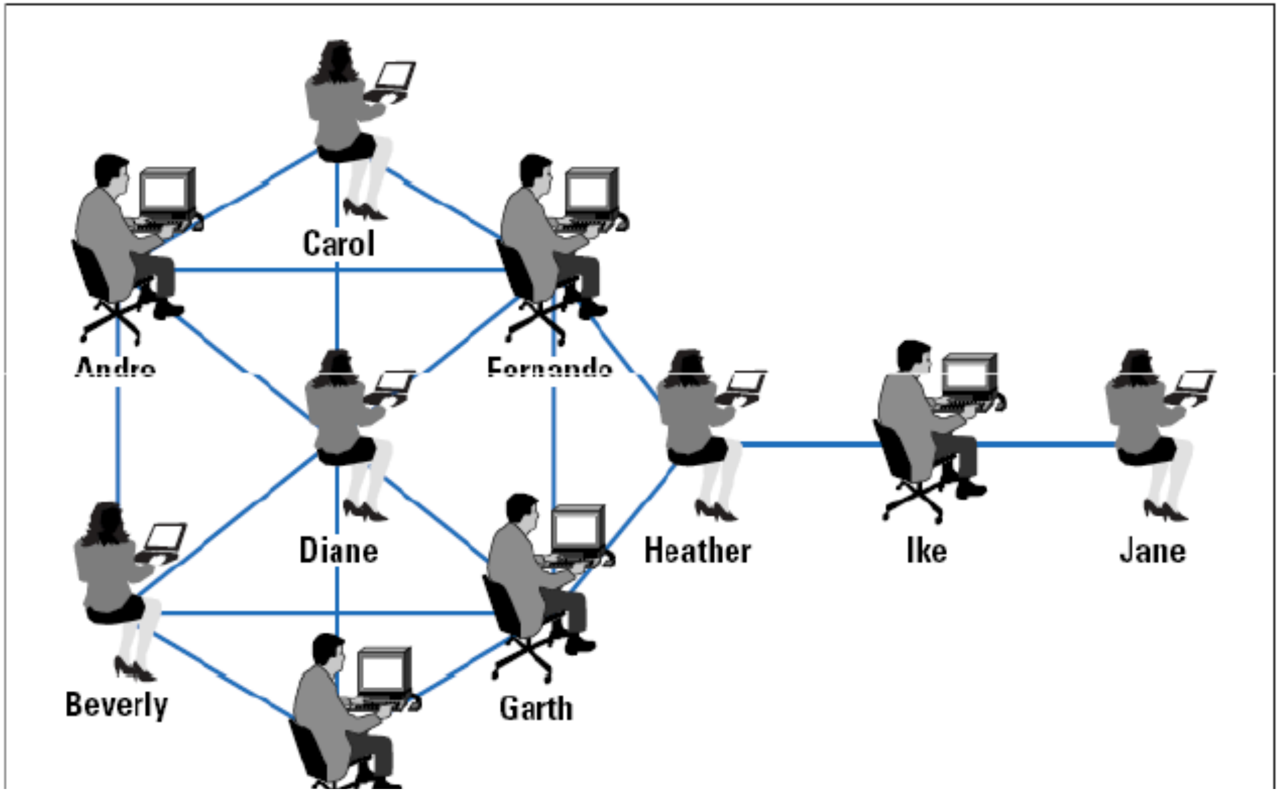
Del.icio.us

User

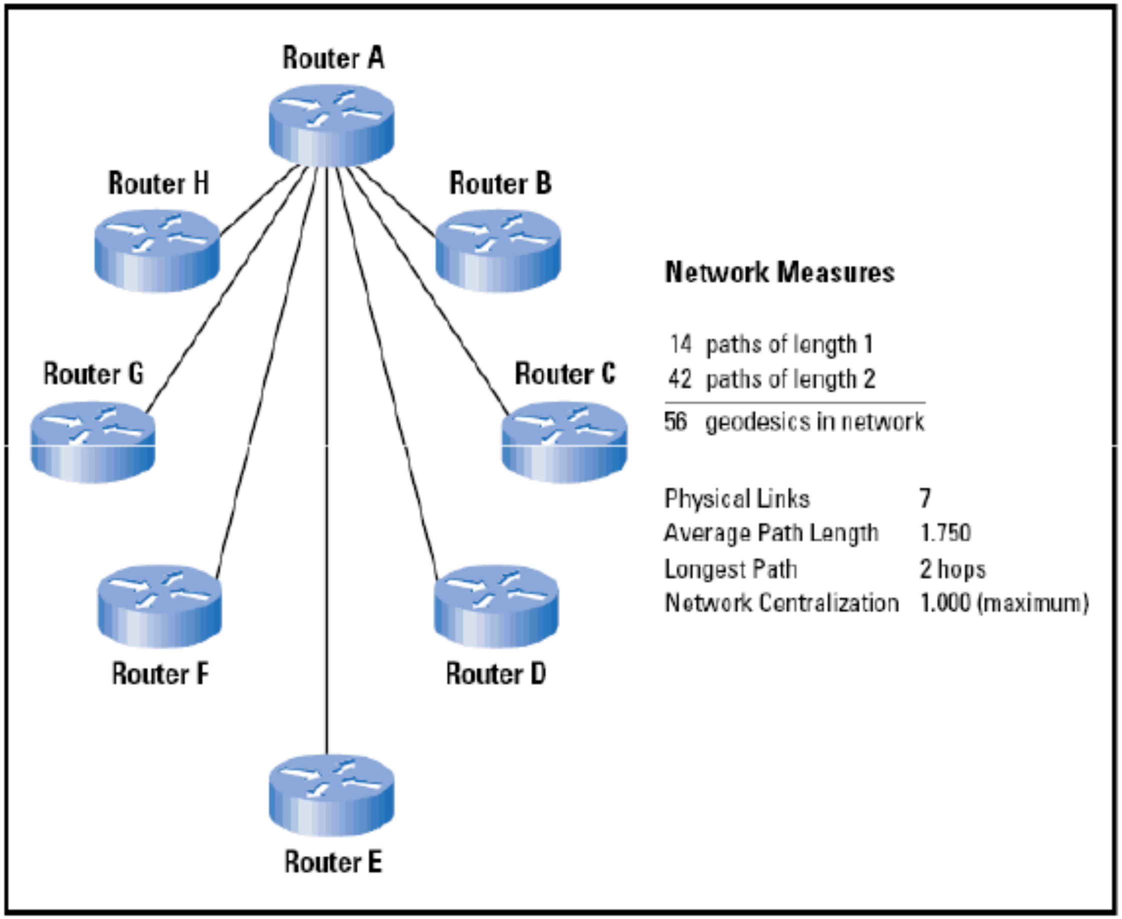
URL

人物关系网＝个体＋关系

Figure 1: Human Network



计算机网络（星型）



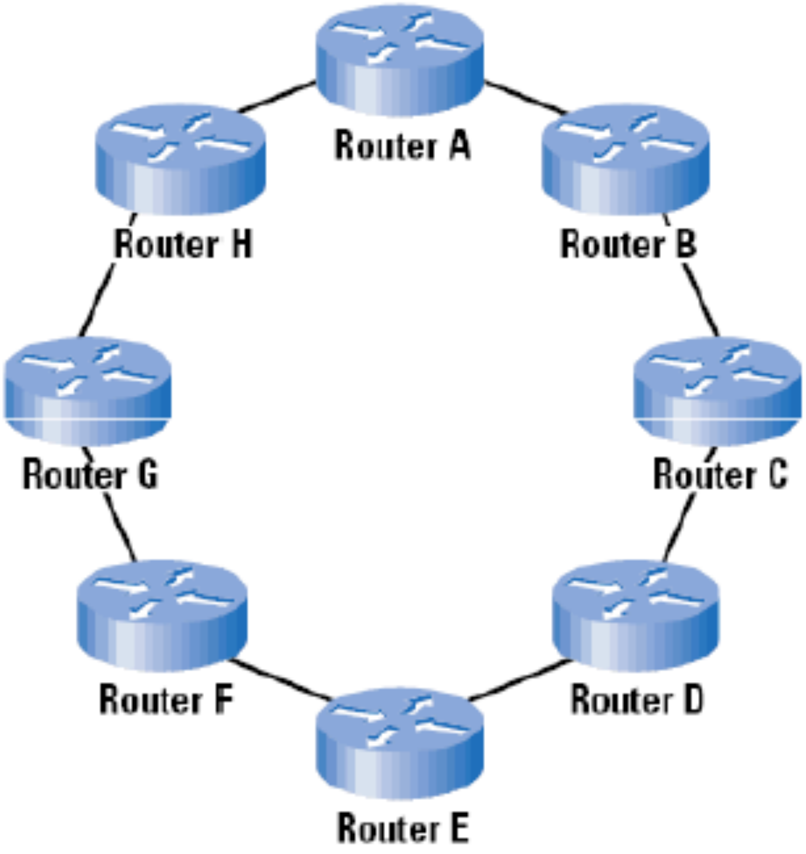
Network Measures

14 paths of length 1
42 paths of length 2

56 geodesics in network

Physical Links	7
Average Path Length	1.750
Longest Path	2 hops
Network Centralization	1.000 (maximum)

计算机网络（环形）



Network Measures

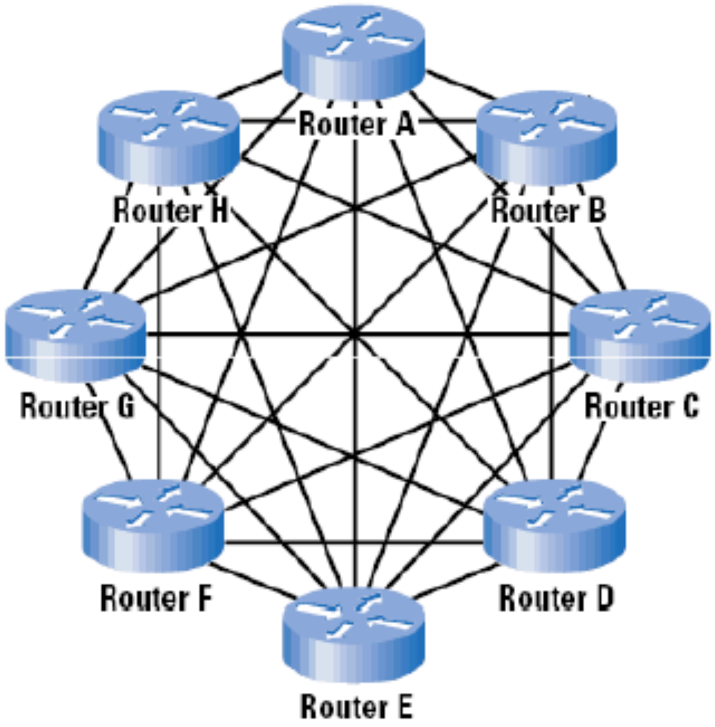
16 paths of length 1
16 paths of length 2
16 paths of length 3
16 paths of length 4

64 geodesics in network

Physical Links	8
Average Path Length	2.500
Longest Path	4 hops
Network Centralization	0.000 (minimum)

计算机网络（全连接）

Figure 4: Routers in Full Mesh Topology



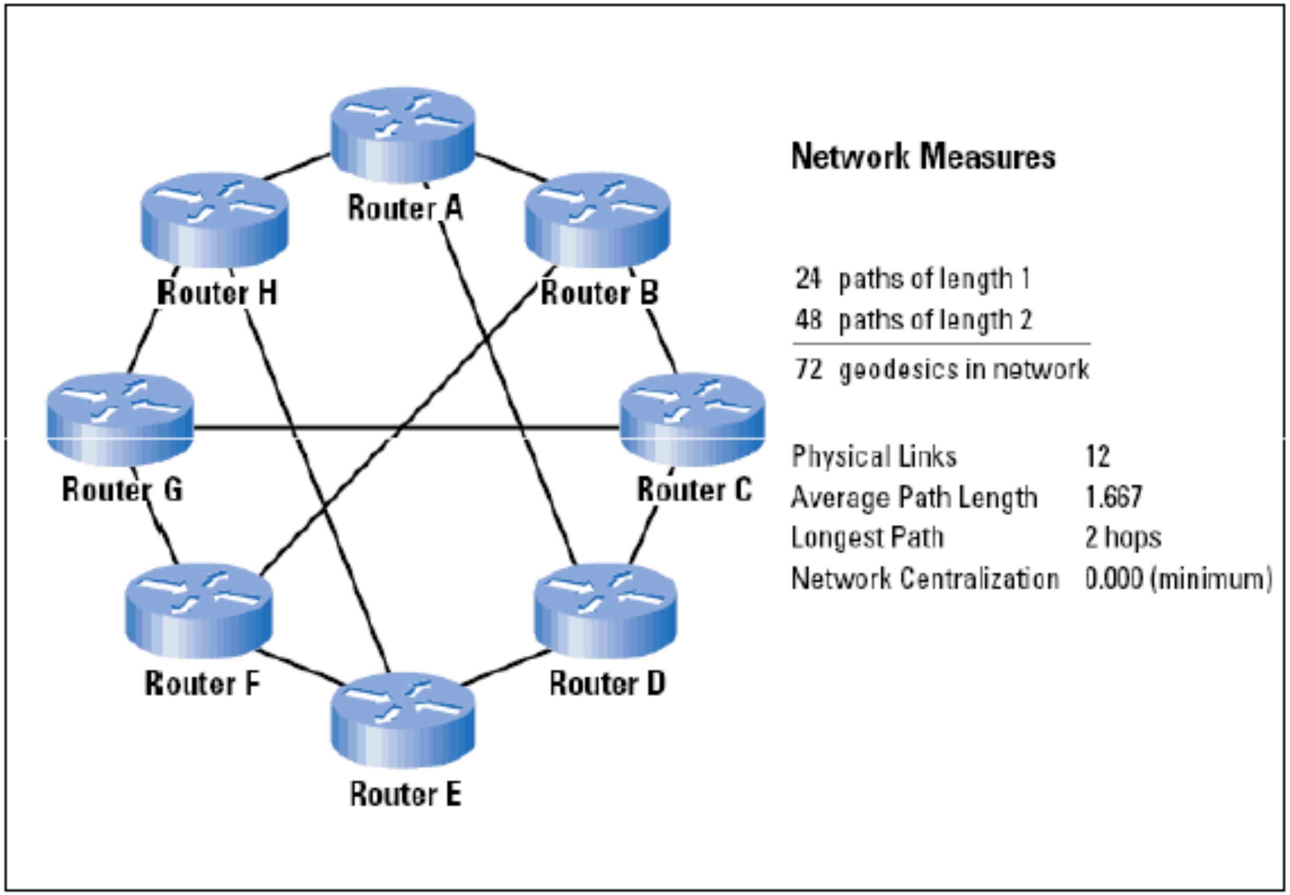
Network Measures

56 paths of length 1
―――
56 geodesics in network

Physical Links	28
Average Path Length	1.000
Longest Path	1 hop
Network Centralization	0.000 (minimum)

计算机网络（部分连接）

Figure 5: Routers in Partial Mesh Topology



Network Measures

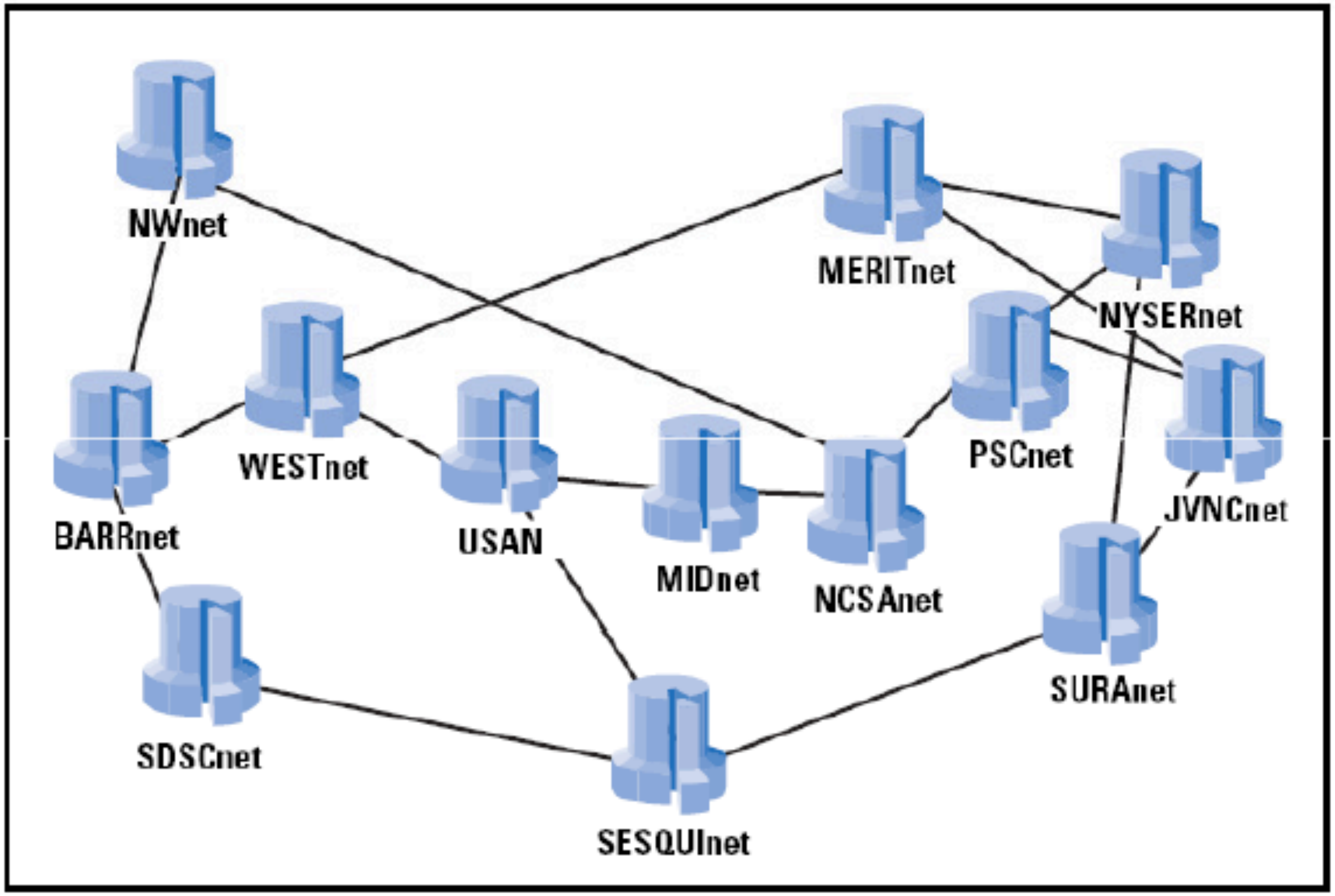
24 paths of length 1
48 paths of length 2

72 geodesics in network

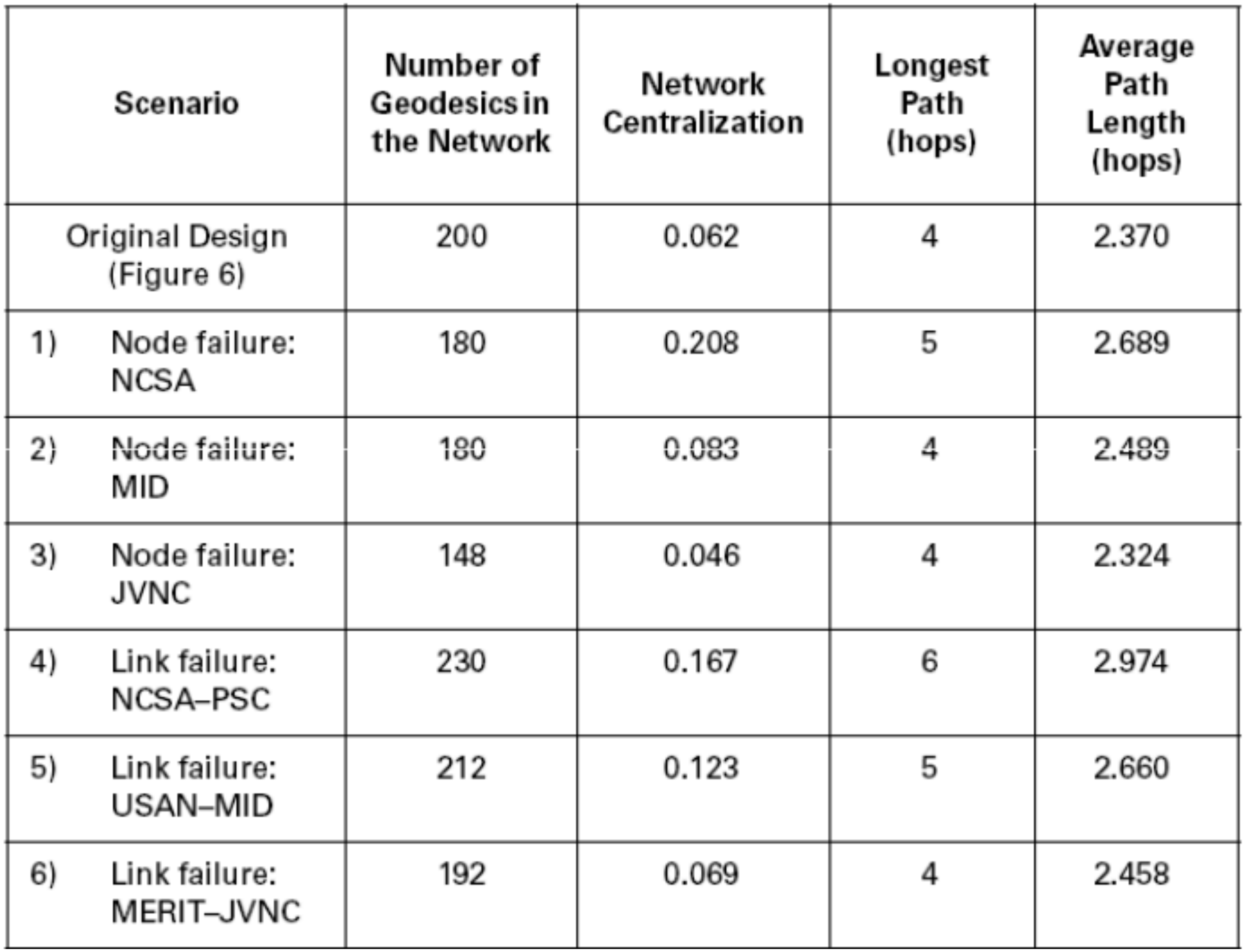
Physical Links	12
Average Path Length	1.667
Longest Path	2 hops
Network Centralization	0.000 (minimum)

1989年NSFnet

igure 6: NSFnet in 1989

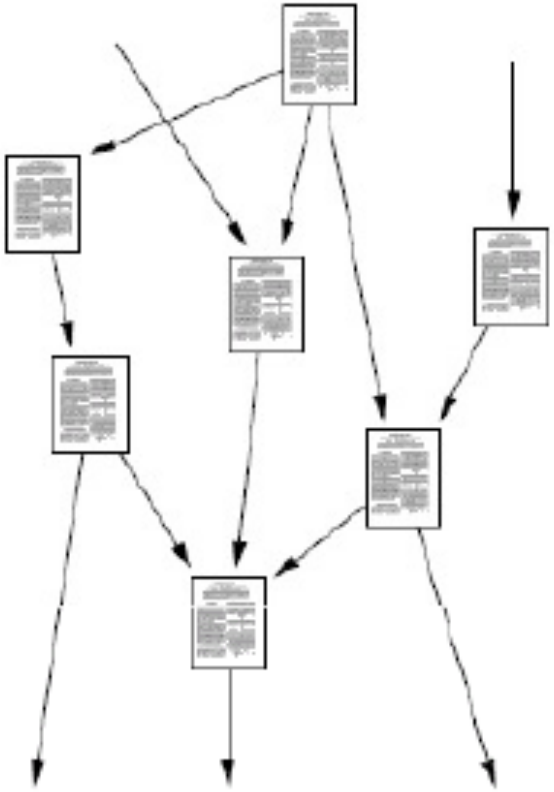


1989年NSFnet的连路和节点失效率

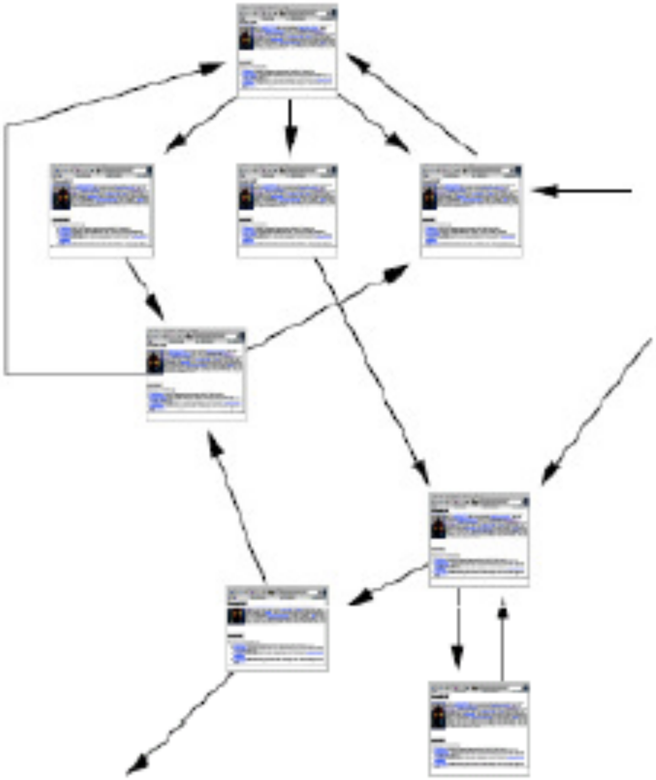


Scenario	Number of Geodesics in the Network	Network Centralization	Longest Path (hops)	Average Path Length (hops)
Original Design (Figure 6)	200	0.062	4	2.370
1) Node failure: NCSA	180	0.208	5	2.689
2) Node failure: MID	180	0.083	4	2.489
3) Node failure: JVNC	148	0.046	4	2.324
4) Link failure: NCSA–PSC	230	0.167	6	2.974
5) Link failure: USAN–MID	212	0.123	5	2.660
6) Link failure: MERIT–JVNC	192	0.069	4	2.458

文章引用与网页连接

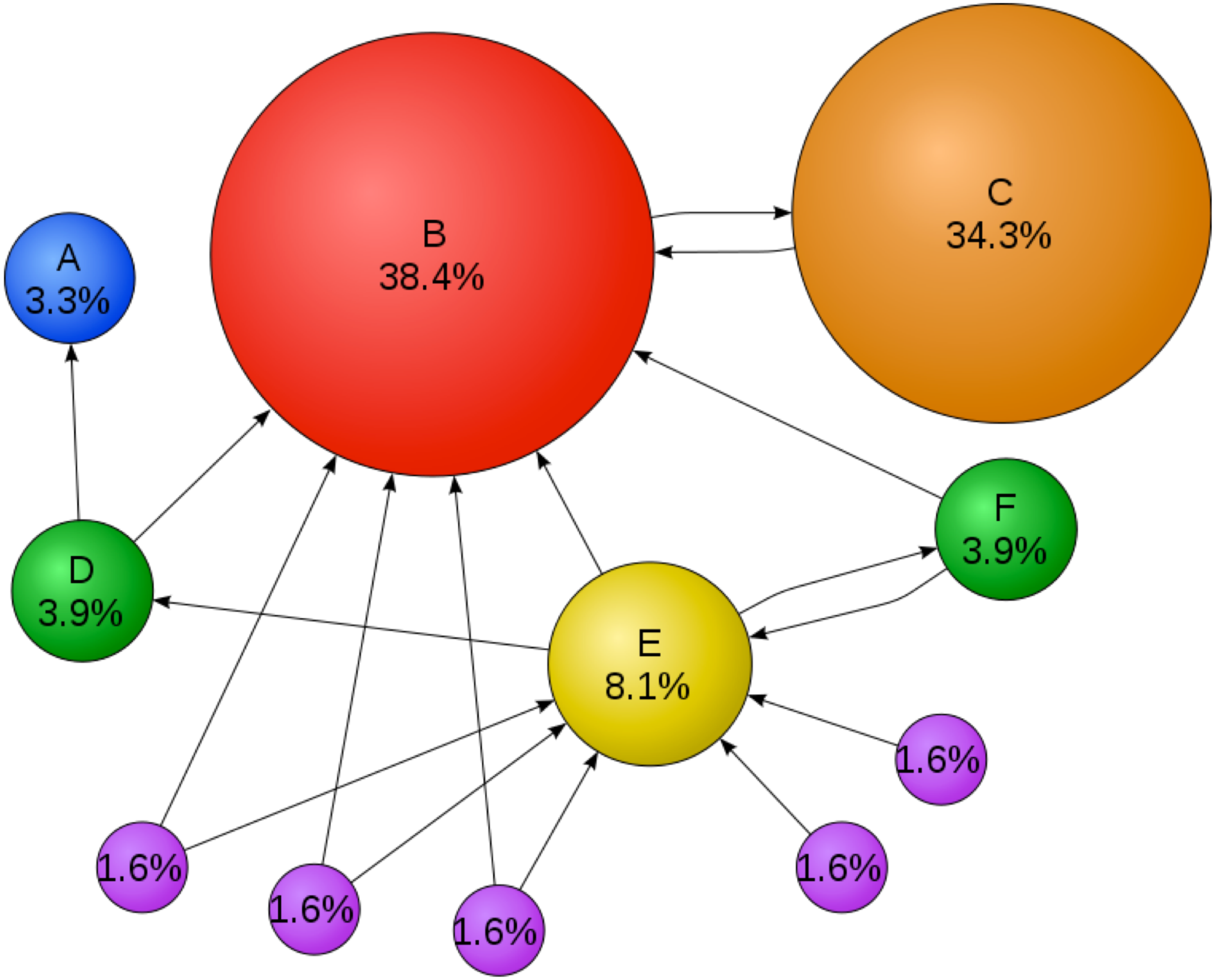


citation network



World–Wide Web

搜索引擎的图解释



其他网络建模

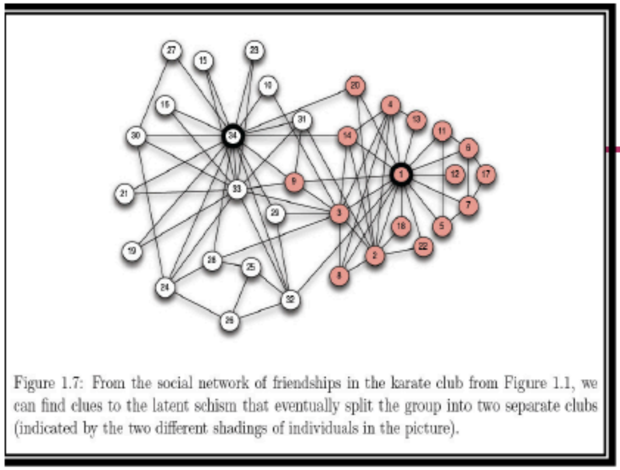


Figure 1.7: From the social network of friendships in the karate club from Figure 1.1, we can find clues to the latent schism that eventually split the group into two separate clubs (indicated by the two different shadings of individuals in the picture).

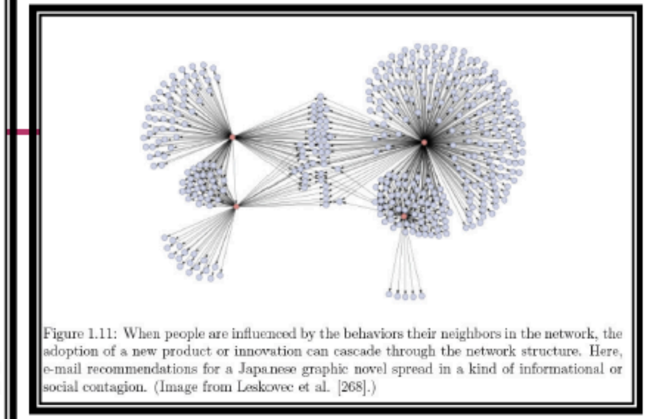


Figure 1.11: When people are influenced by the behaviors their neighbors in the network, the adoption of a new product or innovation can cascade through the network structure. Here, e-mail recommendations for a Japanese graphic novel spread in a kind of informational or social contagion. (Image from Leskovec et al. [268].)

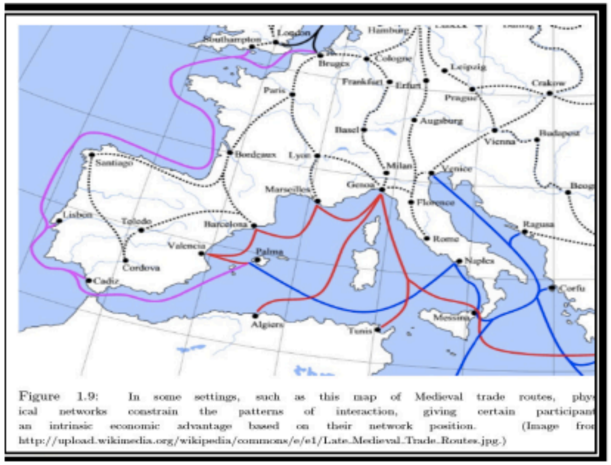
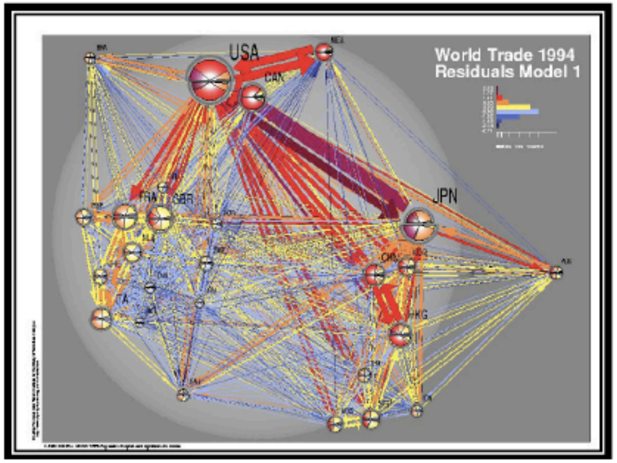
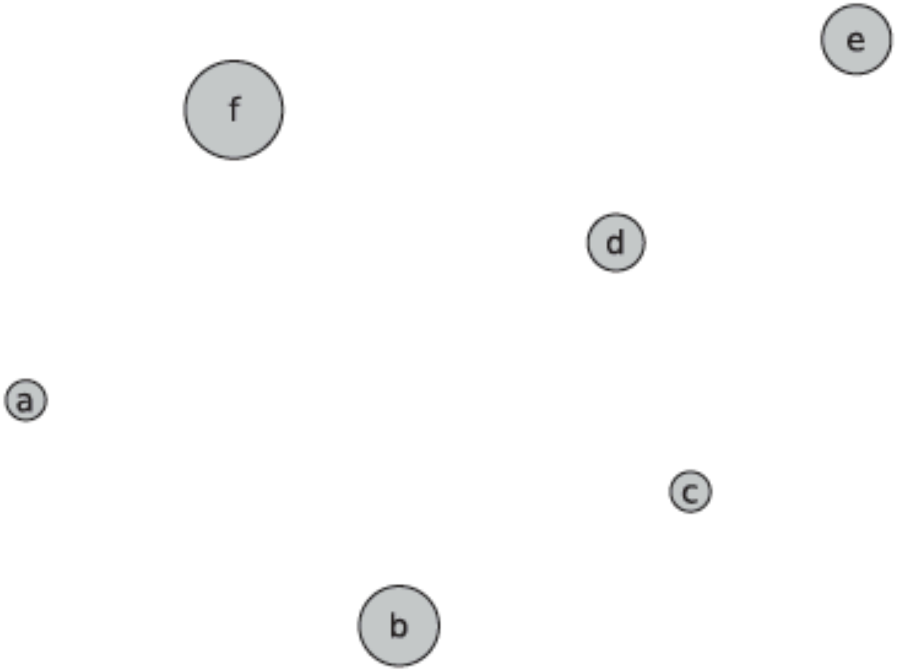


Figure 1.9: In some settings, such as this map of Medieval trade routes, physical networks constrain the patterns of interaction, giving certain participant an intrinsic economic advantage based on their network position. (Image fro http://upload.wikimedia.org/wikipedia/commons/e/e1/Late_Medieval_Trade_Routes.jpg.)



网络实体变量赋值

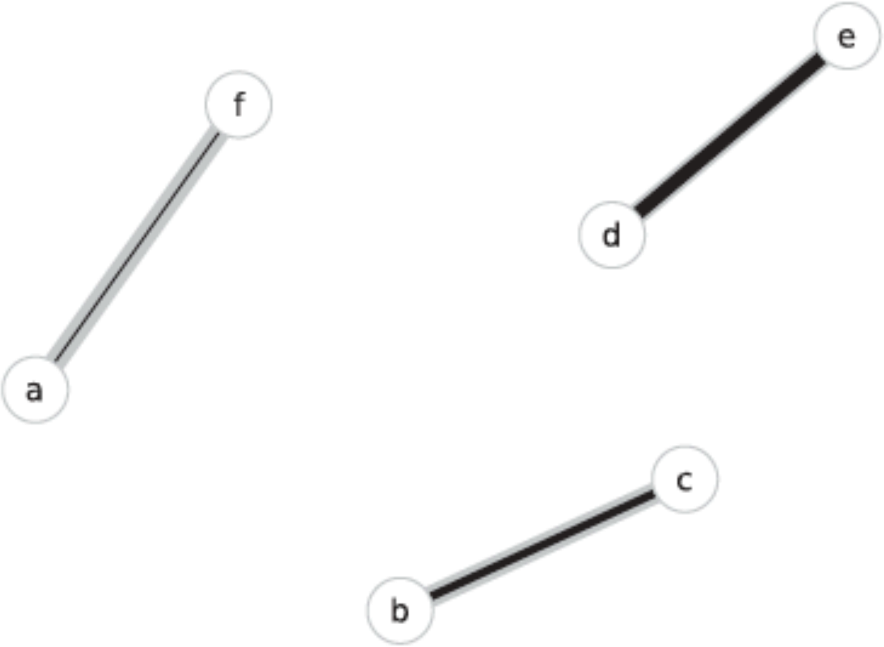


A	x
a	1
b	4
c	1
d	2
e	3
f	6

(a) standard table: variables in columns indexed with unrelated entities

Data formats distinguished by the structure of the domain.

网络实体关系赋值

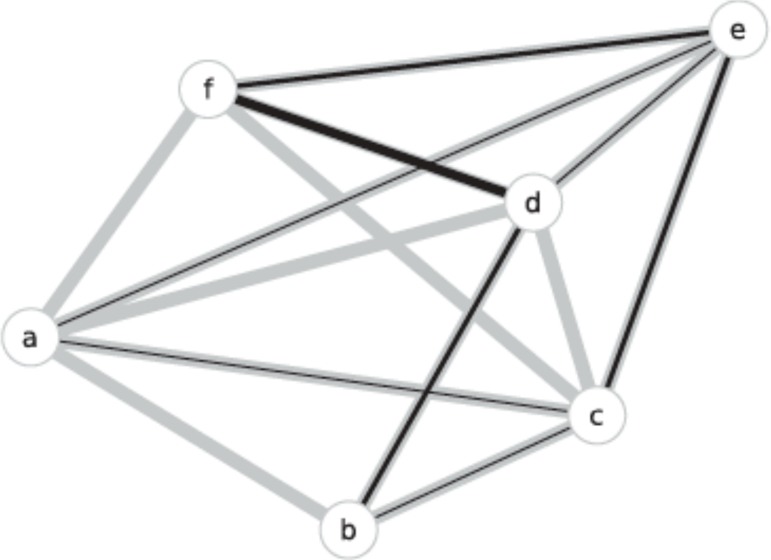


D	x
(a,f)	1
(d,e)	5
(b,c)	3

(b) dyadic: variables in columns indexed with unrelated pairs of entities

Data formats distinguished by the structure of the domain.

网络的储存表示



D	x
(a,b)	0
(a,c)	1
(a,d)	0
(a,e)	1
(a,f)	0
(b,c)	1
(b,d)	2
(c,d)	0

⋮

$x(D)$	a	b	c	d	e	f
a	·	0	1	0	1	0
b	0	·	1	2	·	·
c	1	1	·	0	2	0
d	0	2	0	·	1	4
e	1	·	2	1	·	2
f	0	·	0	4	2	·

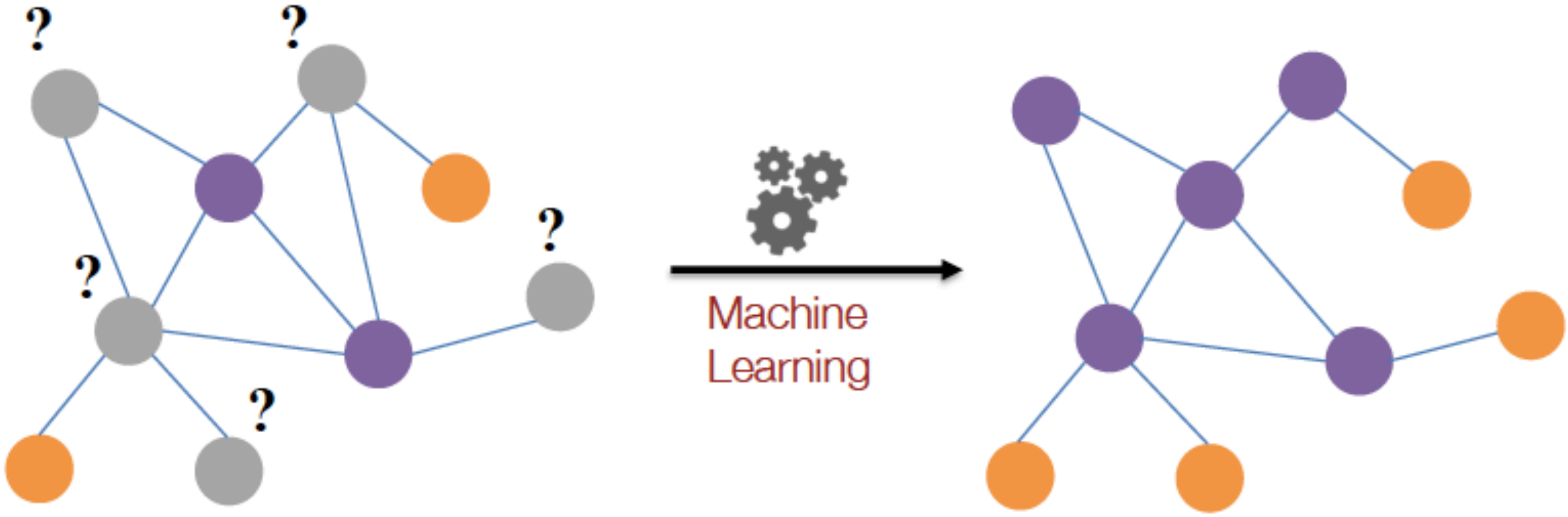
(c) network: variables in columns indexed with incident pairs of entities, or in matrices

Data formats distinguished by the structure of the domain.

网络建模推理演算中的机器学习任务

- Node classification
  - Predict a type of a given node
- Link prediction
  - Predict whether two nodes are linked
- Community detection
  - Identify densely linked clusters of nodes
- Network similarity
  - How similar are two (sub)networks

网络的节点分类任务

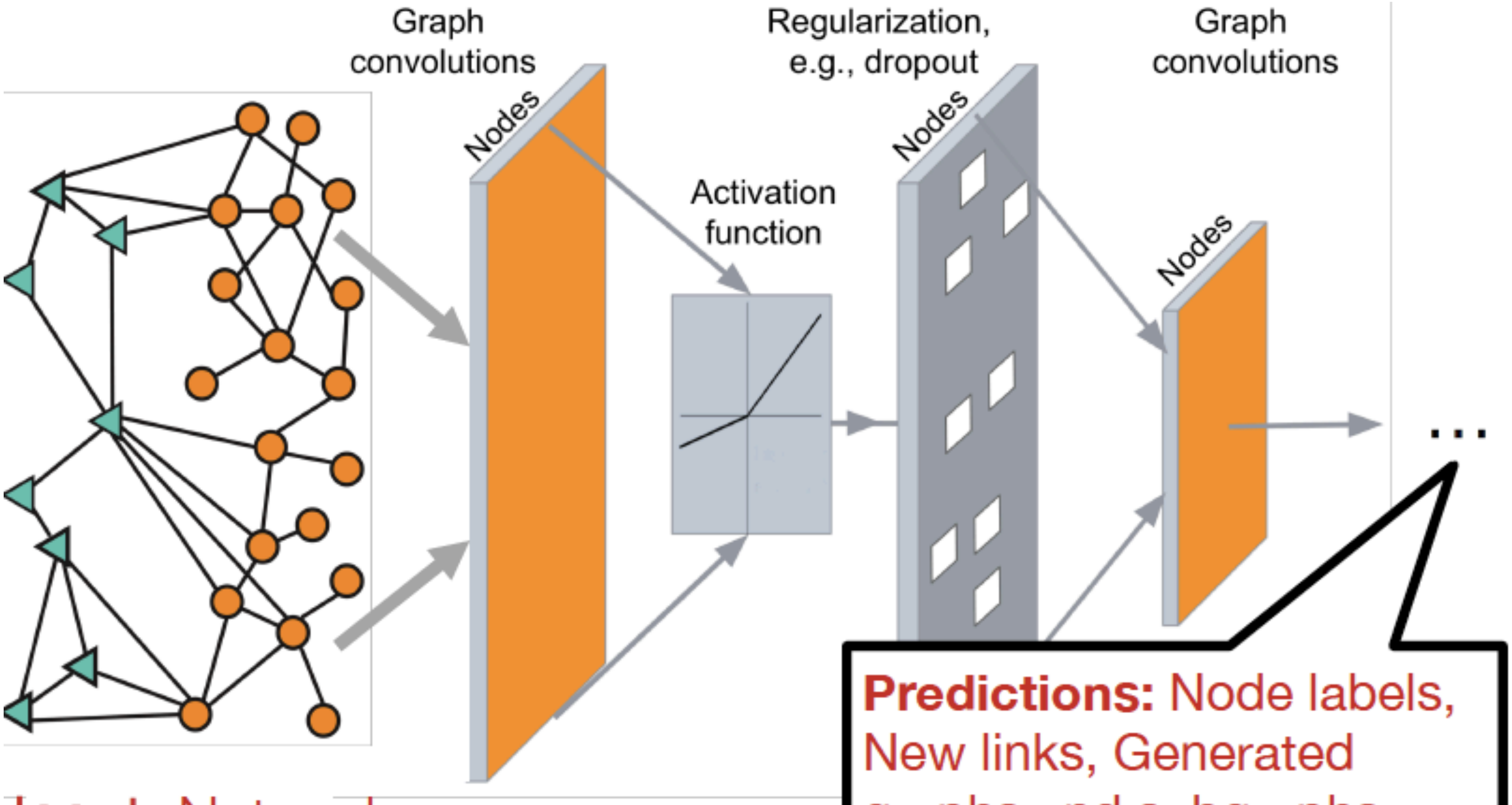


Example: Node Classification

Many possible ways to create node features:

- Node degree, PageRank score, Motifs, Degree of neighbors, Clustering, …

深度学习与图的预测任务

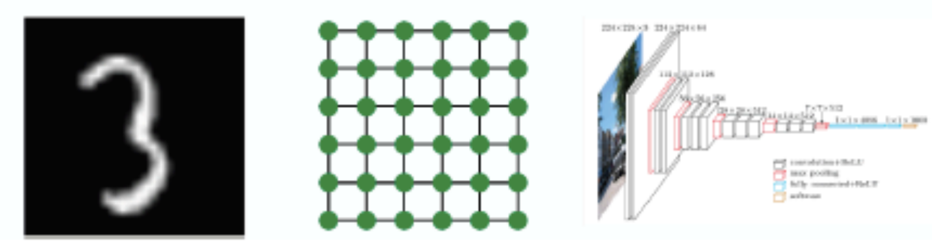


Input: Network

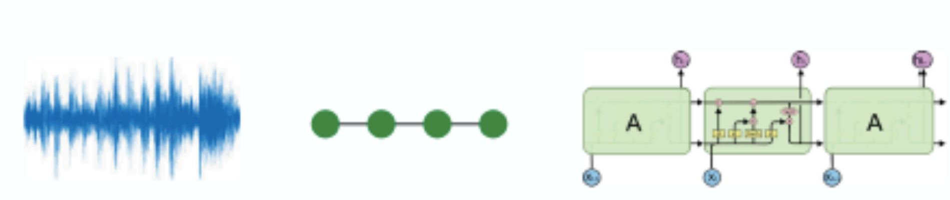
Predictions: Node labels, New links, Generated graphs and subgraphs

深度学习在图预测任务的困难

- Modern deep learning toolbox is designed for simple sequences or grids
  - CNNs for fixed-size images/grids….



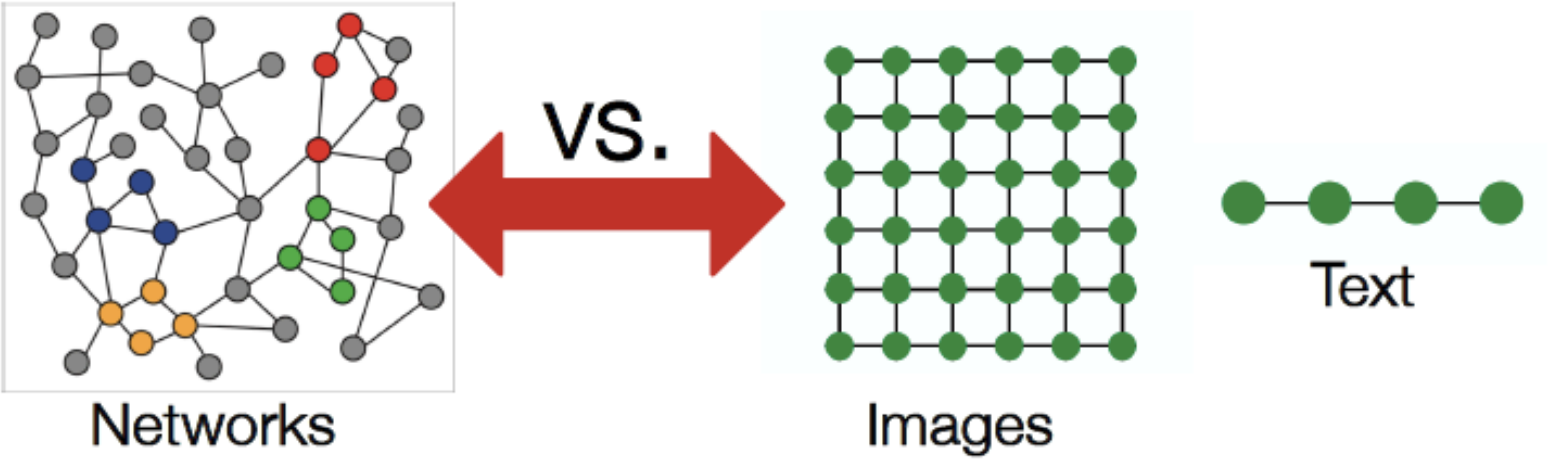
  - RNNs or word2vec for text/sequences…



深度学习在图预测任务的困难

But networks are far more complex!

- Arbitrary size and complex topological structure (i.e., no spatial locality like grids)



- No fixed node ordering or reference point
- Often dynamic and have multimodal features

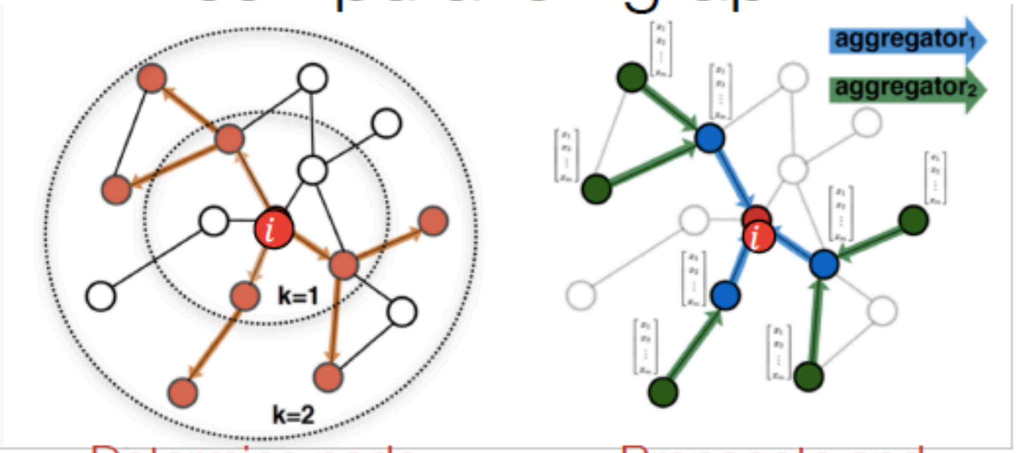
图的建模方法

We have a graph G:

- V is the vertex set
- $\boldsymbol{A}$ is the (binary) adjacency matrix
- $\boldsymbol{X} \in \mathbb{R}^{m \times |V|}$ is a matrix of node features
  - Meaningful node features:
    - Social networks: User profile
    - Biological networks: Gene expression profiles, gene functional information

面向图的神经网络（1）

Idea: Node's neighborhood defines a computation graph



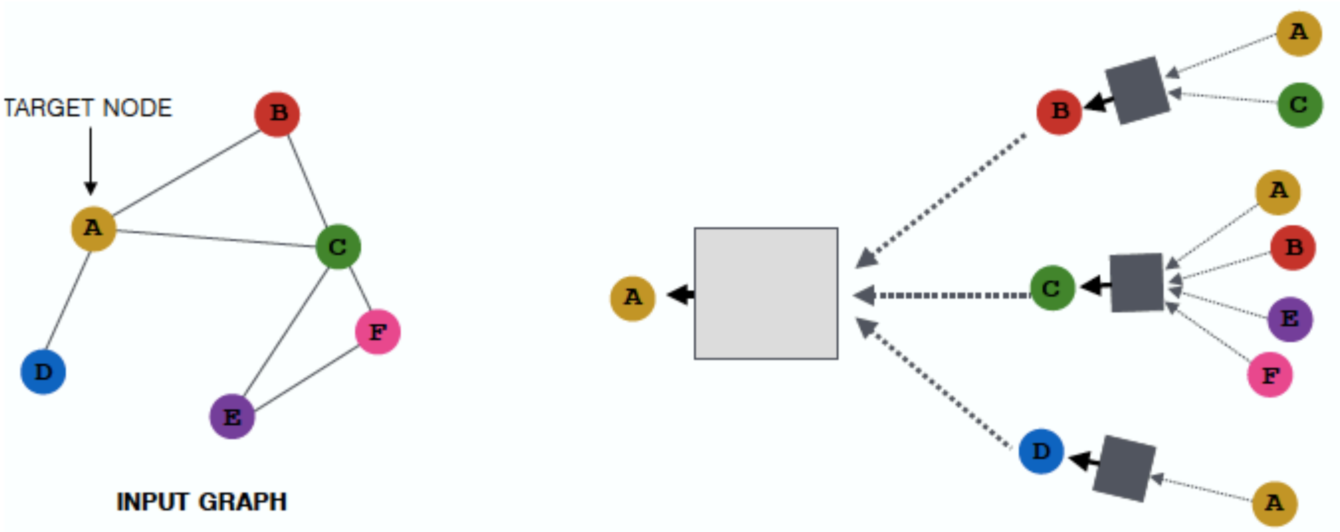
Determine node computation graph

Propagate and transform information

Learn how to propagate information across the graph to compute node features

The Graph Neural Network Model. Scarselli et al. *IEEE Transactions on Neural Networks* 2005
Semi-Supervised Classification with Graph Convolutional Networks. T. N. Kipf, M. Welling, ICLR 2017

面向图的神经网络（2）



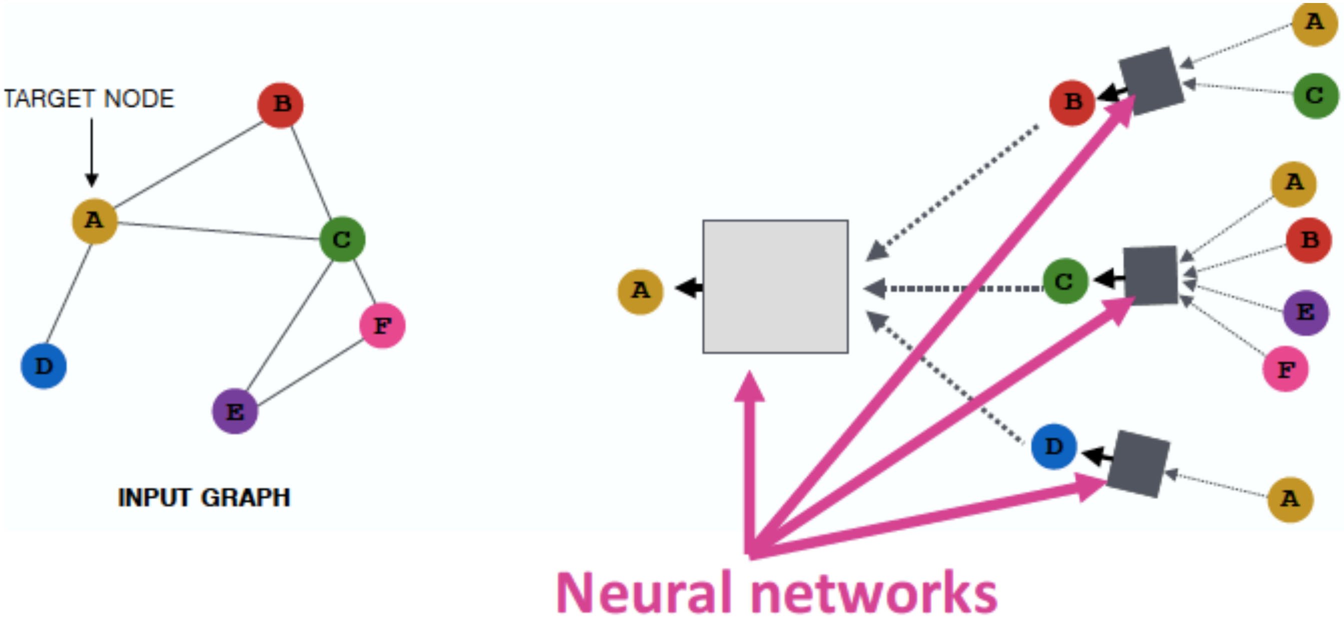
Each node defines a computation graph

- Each edge in this graph is a transformation/aggregation function

Inductive Representation Learning on Large Graphs. W. Hamilton, R. Ying, J. Leskovec. NIPS, 2017.

中国科学院大学

面向图的神经网络（3）



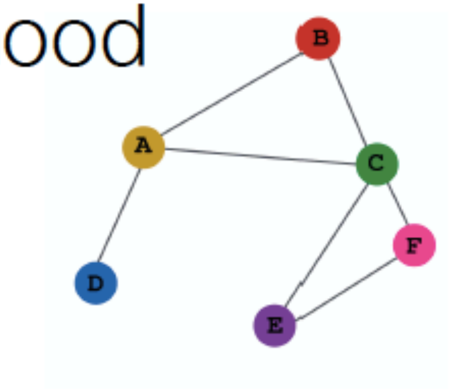
Intuition: Nodes aggregate information from their neighbors using neural networks

Inductive Representation Learning on Large Graphs. W. Hamilton, R. Ying, J. Leskovec. NIPS, 2017.

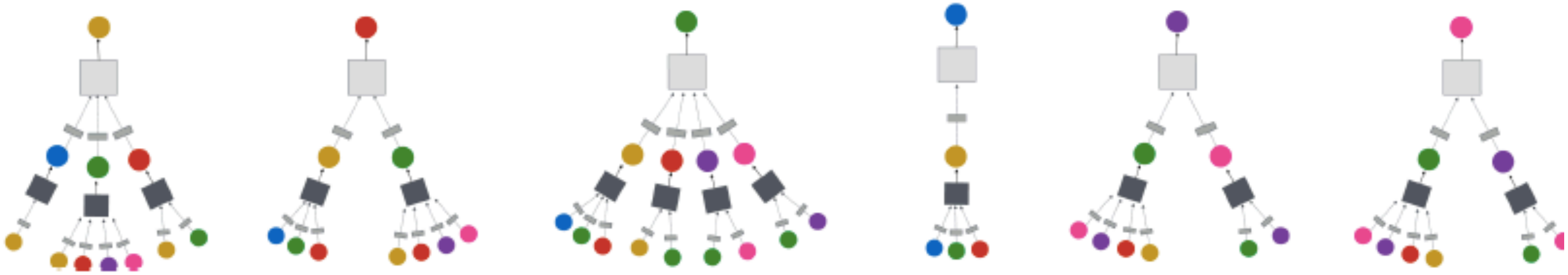
中国科学院大学

邻近节点聚合的算法思路

Intuition: Network neighborhood defines a computation graph



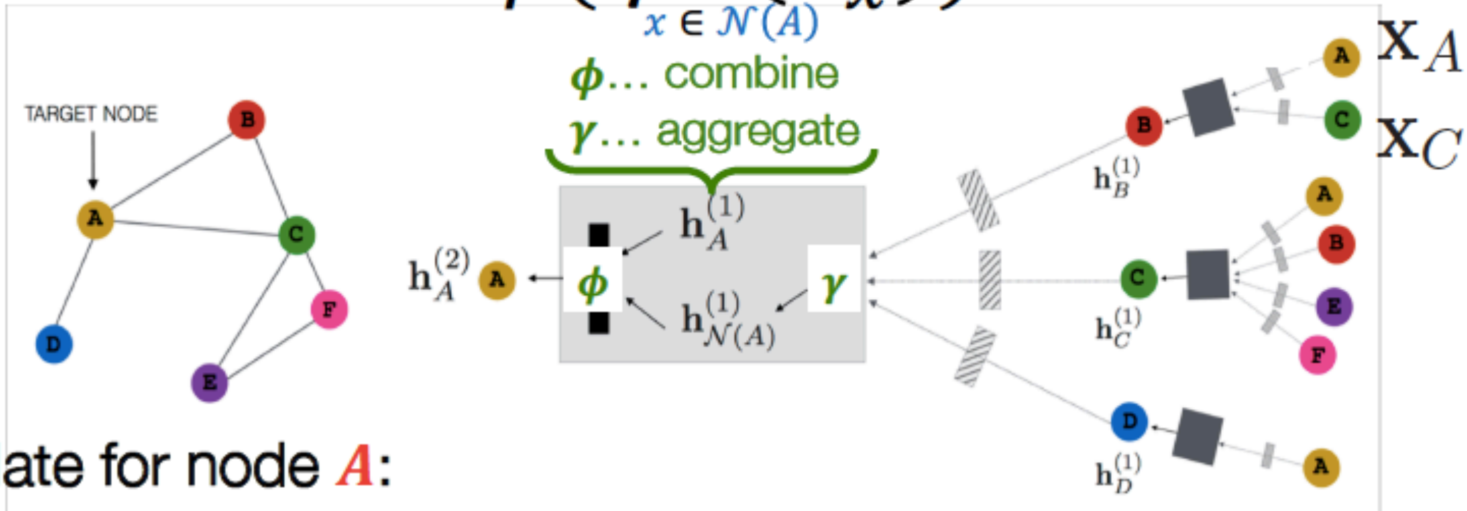
Every node defines a computation graph based on its neighborhood!



Can be viewed as learning a generic linear combination of graph low-pass and high-pass operators

Graph SAGE 方法

$$\phi(\gamma_{x\in\mathcal{N}(A)}(h_x))$$



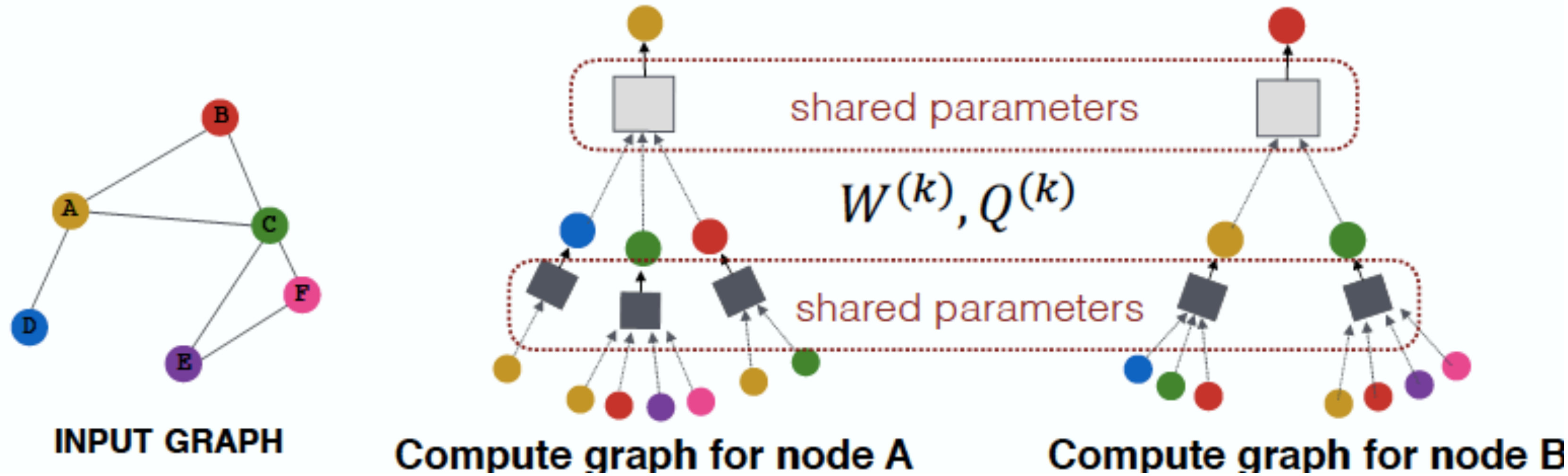
Update for node A:

$$\underbrace{h_A^{(k+1)}}_{\text{$k+1^{st}$ level embedding of node A}} = \sigma\left(\underbrace{W^{(k)}h_A^{(k)}}_{\text{Transform A's own embedding from level k}}, \gamma_{x\in\mathcal{N}(A)}\left(\underbrace{\sigma(Q^{(k)}h_x^{(k)})}_{\text{Transform and aggregate embeddings of neighbors n}}\right)\right)$$

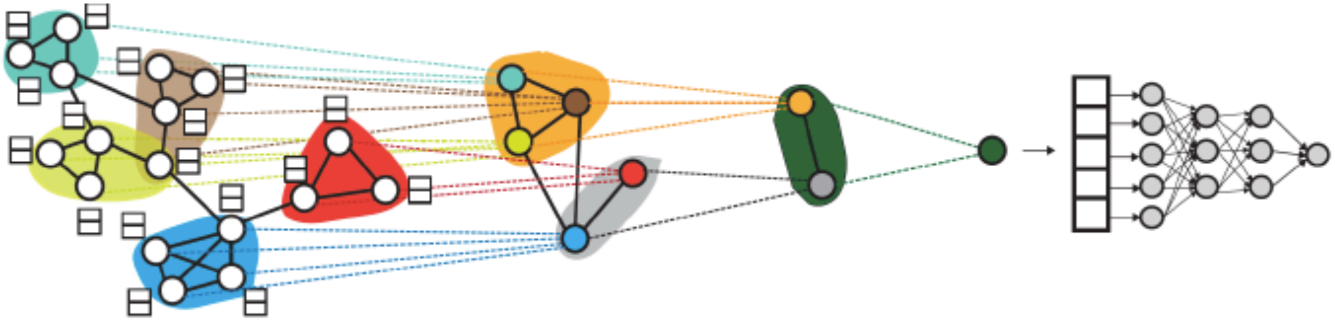
- $h_A^{(0)}$ = attributes X_A of node A

Graph SAGE 训练

- Aggregation parameters are shared for all nodes
- The number of model parameters is sublinear in |V|
- Can use different loss functions:
  - Classification/Regression: $\mathcal{L}(h_A) = ||y_A - f(h_A)||^2$
  - Pairwise Loss: $\mathcal{L}(h_A, h_B) = \max(0, 1 - dist(h_A, h_B))$



Graph SAGE的 GNNs池化



Don’t just embed individual nodes. Embed the entire graph.

Problem: Learn how to hierarchical pool the nodes to embed the entire graph

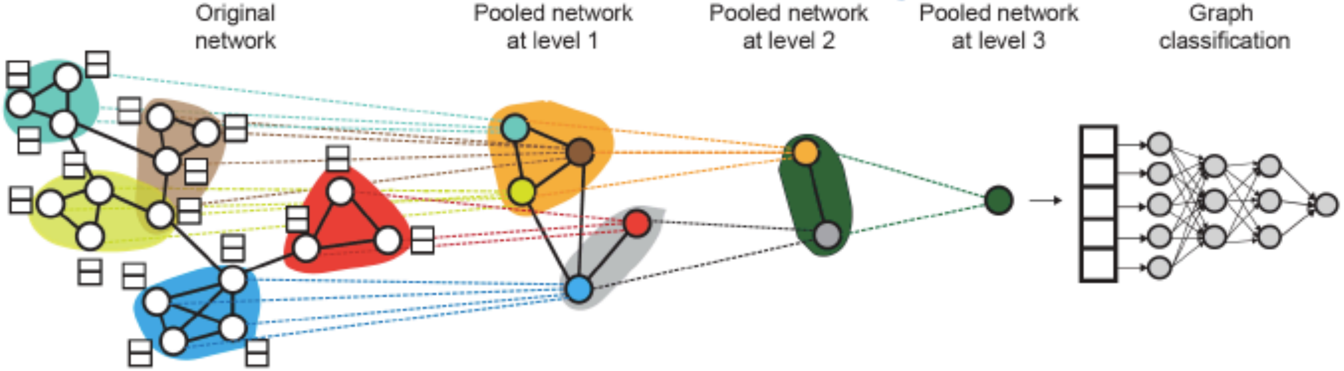
Our solution: **DiffPool**

- Learns hierarchical pooling strategy
- Sets of nodes are pooled hierarchically
- Soft assignment of nodes to next-level nodes

Hierarchical Graph Representation Learning with Differentiable Pooling. R. Ying, et al. NeurIPS 2018.

Graph SAGE的 DiffPool 架构

A different GNN is learned at every level of abstraction


Original network
Pooled network at level 1
Pooled network at level 2
Pooled network at level 3
Graph classification

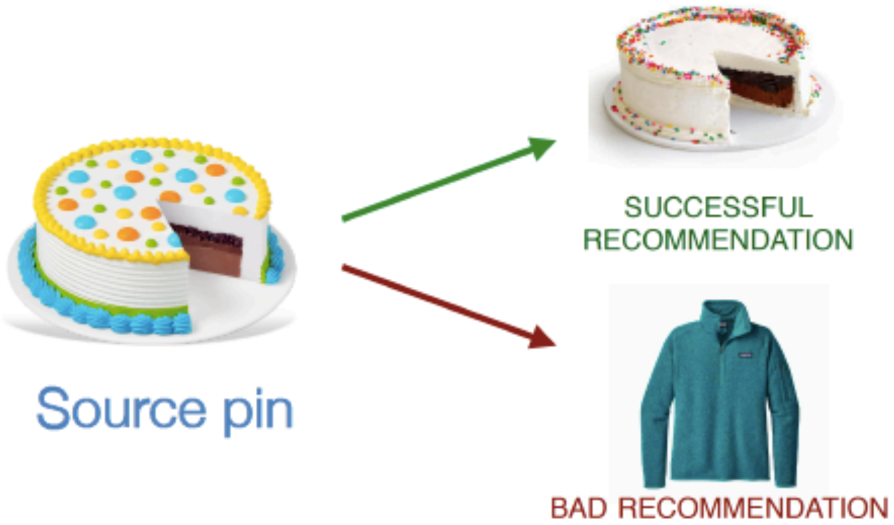
Our approach: Use two sets of GNNs

- GNN1 to learn how to pool the network
  - Learn cluster assignment matrix
- GNN2 to learn the node embeddings

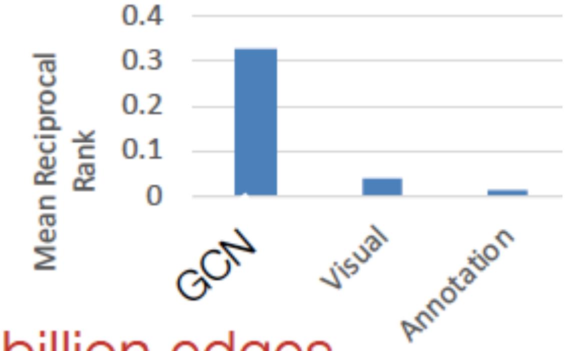
Hierarchical Graph Representation Learning with Differentiable Pooling. R. Ying, et al. NeurIPS 2018.

推荐图片的相关性

Task: Recommend related pins



Task: Learn node embeddings z_i s.t. $d(z_{cake1}, z_{cake2}) < d(z_{cake1}, z_{sweater})$



- **Challenges:**
  - Massive size: 3 billion nodes, 20 billion edges
  - Heterogeneous data: Rich image and text features

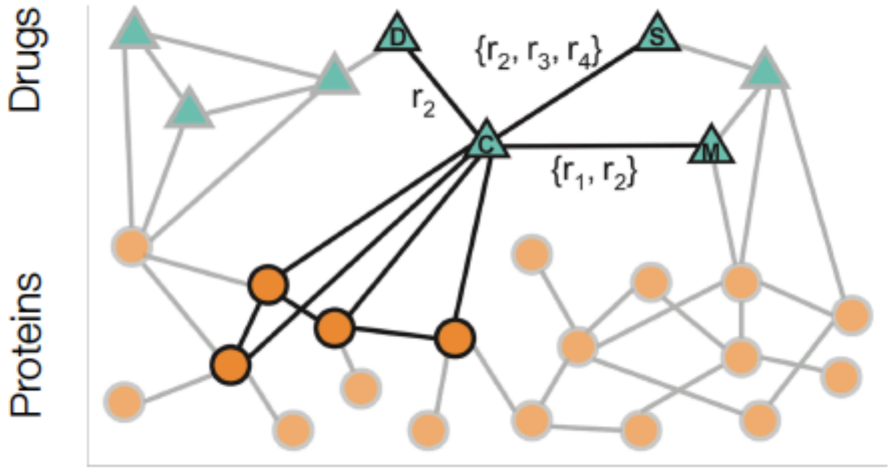
Graph Convolutional Neural Networks for Web-Scale Recommender Systems. R. Ying, et al. KDD. 2018.

预测药物的副作用

- **Task: Given a pair of drugs predict adverse side effects**

  46% of people ages 70-79 take >5 drugs

- Link prediction on a multimodal graph



36% improvement in AP@50 over state of the art

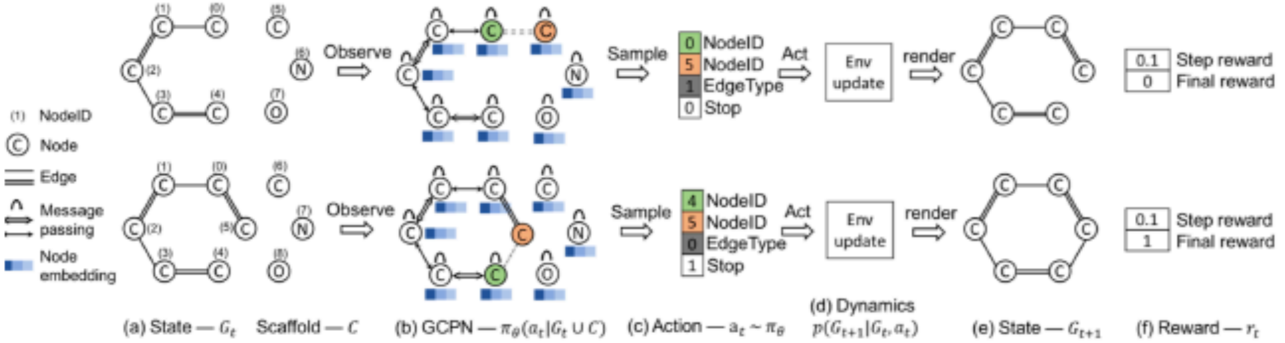
Modeling Polypharmacy Side Effects with Graph Convolutional Networks. M. Zitnik, et al. Bioinformatics, 2018.

目标分子的生成

Goal: Generate molecules that optimize a given property (Quant. energy, solubility)

Solution: Combination of

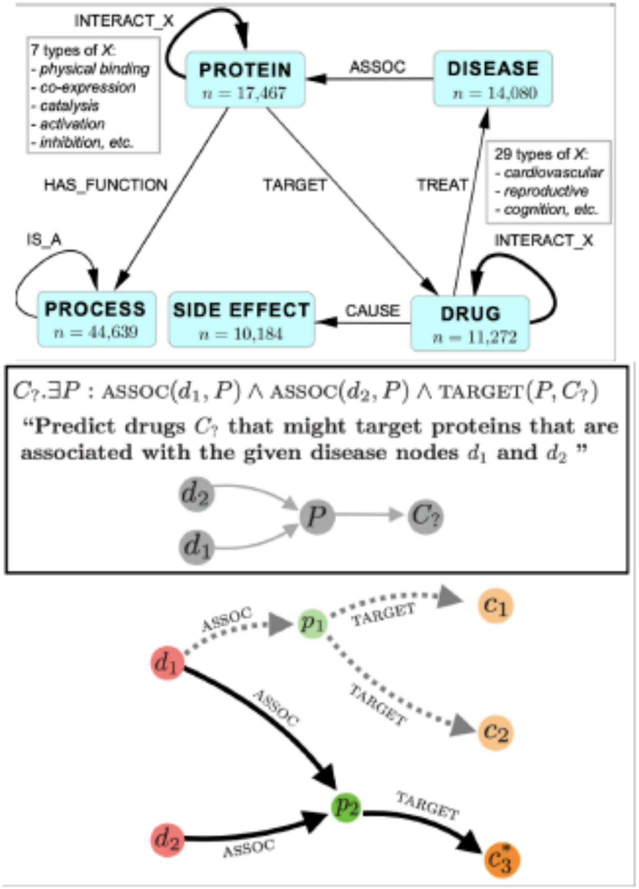
- Graph representation learning
- Adversarial training
- Reinforcement learning



(a) State — G_t Scaffold — C (b) GCPN — $\pi_\theta(a_t|G_t \cup C)$ (c) Action — $a_t \sim \pi_\theta$ (d) Dynamics $p(G_{t+1}|G_t, a_t)$ (e) State — G_{t+1} (f) Reward — r_t

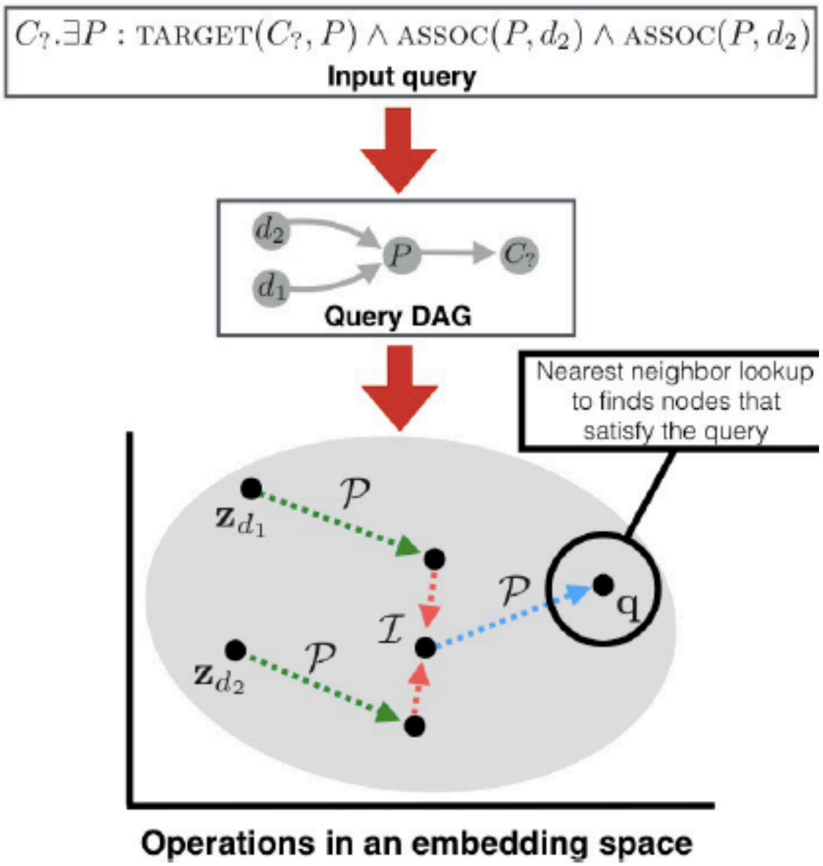
Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. J. You, et al., NeurIPS 2018.

知识图谱（Knowledge Graph)



$C_?.\exists P : \textsc{assoc}(d_1, P) \wedge \textsc{assoc}(d_2, P) \wedge \textsc{target}(P, C_?)$

"Predict drugs $C_?$ that might target proteins that are associated with the given disease nodes d_1 and d_2 "



$C_?.\exists P : \textsc{target}(C_?, P) \wedge \textsc{assoc}(P, d_2) \wedge \textsc{assoc}(P, d_2)$

Input query

Operations in an embedding space

Embedding Logical Queries on Knowledge Graphs. W. Hamilton, et al. NeurIPS, 2018.

Graph SAGE 总结

- Graph Convolution Networks
  - Generalize beyond simple convolutions
- Fuses node features & graph info
  - State-of-the-art accuracy for node classification and link prediction.
- Model size independent of graph size; can scale to billions of nodes
  - Largest embedding to date (3B nodes, 17B edges)
- Leads to significant performance gains

结论

Results from the past 2-3 years have shown:

- Representation learning paradigm can be extended to graphs
- No feature engineering necessary
- Can effectively combine node attribute data with the network information
- State-of-the-art results in a number of domains/tasks
- Use end-to-end training instead of multi-stage approaches for better performance

References

- Tutorial on Representation Learning on Networks at WWW 2018 http://snap.stanford.edu/proj/embeddings-www/
- Inductive Representation Learning on Large Graphs.
  W. Hamilton, R. Ying, J. Leskovec. NIPS 2017.
- Representation Learning on Graphs: Methods and Applications. W. Hamilton, R. Ying, J. Leskovec. IEEE Data Engineering Bulletin, 2017.
- Graph Convolutional Neural Networks for Web-Scale Recommender Systems. R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, J. Leskovec. KDD, 2018.
- Modeling Polypharmacy Side Effects with Graph Convolutional Networks. M. Zitnik, M. Agrawal, J. Leskovec. Bioinformatics, 2018.
- Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. J. You, B. Liu, R. Ying, V. Pande, J. Leskovec, NeurIPS 2018.
- Embedding Logical Queries on Knowledge Graphs. W. Hamilton, P. Bajaj, M. Zitnik, D. Jurafsky, J. Leskovec. NeuIPS, 2018.
- How Powerful are Graph Neural Networks? K. Xu, W. Hu, J. Leskovec, S. Jegelka. ICLR 2019.

- Code:
  - http://snap.stanford.edu/graphsage
  - http://snap.stanford.edu/decagon/
  - https://github.com/bowenliu16/rl_graph_generation
  - https://github.com/williamleif/graphqembed
  - https://github.com/snap-stanford/GraphRNN