E6885 Network Science Lecture 1:
Overview of Network Science

Ching-Yung Lin, Dept. of Electrical Engineering, Columbia University

September 12th, 2011

Course Structure

Class Date	Class Number	Topics Covered
09/12/11	1	Overview – Social, Information, and Cognitive Network Analysis
09/19/11	2	Network Representations and Characteristics
09/26/11	3	Network Partitioning and Clustering
10/03/11	4	Network Visualization
10/10/11	5	Network Sampling and Estimation
10/17/11	6	Network Models
10/24/11	7	Network Topology Inference
10/31/11	8	Dynamic Networks
11/14/11	9	Social Influence and Info Diffusion in Networks
11/21/11	10	Information and Knowledge Networks
11/28/11	11	Privacy, Security, and Economy Issues in Networks
12/05/11	12	Behavior Understanding and Cognitive Networks
12/12/11	13	Large-Scale Network Processing System
12/19/11	14	Final Project Presentation

Course Grading

- 4 Homeworks: 50%

- Final Project: 50%

Other Issues

- Office Hours: Monday 6:40pm – 7:10pm (Mudd 1312)
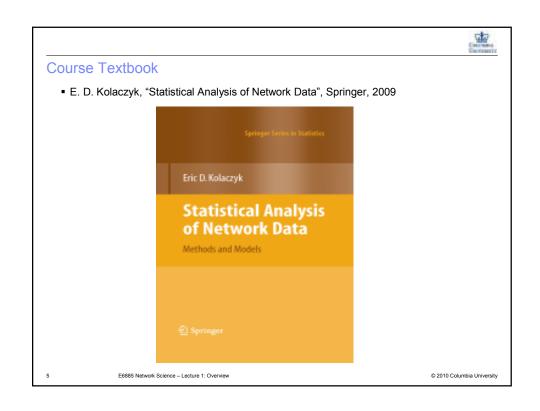
- TA / Grader: TBA
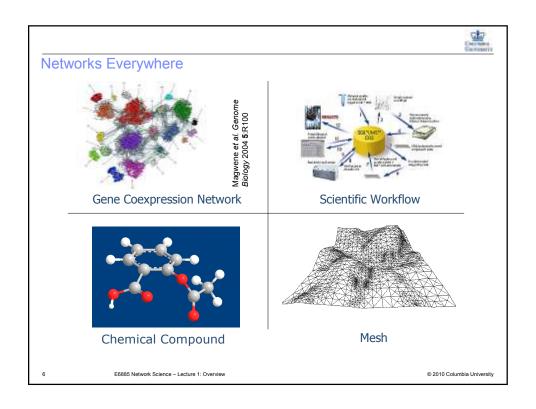
- Contact: cylin@ee.columbia.edu

- Telephone: 914-784-7822

- Twitter: CUNetSci

Course Textbook

- E. D. Kolaczyk, "Statistical Analysis of Network Data", Springer, 2009

Springer Series in Statistics

Eric D. Kolaczyk

Statistical Analysis of Network Data

Methods and Models

Springer

Networks Everywhere

Gene Coexpression Network

Magwene et al. Genome Biology 2004 **5**:R100

Scientific Workflow

Chemical Compound

Mesh

First Reported Network Analysis

The New York Times.

Copyright, 1933, by The New York Times Company.

Entered as Second-Class Matter, Postoffice, New York, N. Y. — NEW YORK, THURSDAY, APRIL 3, 1933. — TWO CENTS

EMOTIONS MAPPED BY NEW GEOGRAPHY

Charts Seek to Portray the Psychological Currents of Human Relationships.

FIRST STUDIES EXHIBITED

Colored Lines Show Likes and Dislikes of Individuals and of Groups.

MANY MISFITS REVEALED

Dr. J. L. Moreno Calculates There Are 10 to 15 Million Isolated Individuals in Nation.

A new science, named psychological geography, which aims to chart the emotional currents, cross-currents and under-currents of human relationships in a community, was introduced here yesterday at the scientific exhibit of the Medical Society of the State of New York, which opens its 127th annual meeting here today at the Waldorf-Astoria.
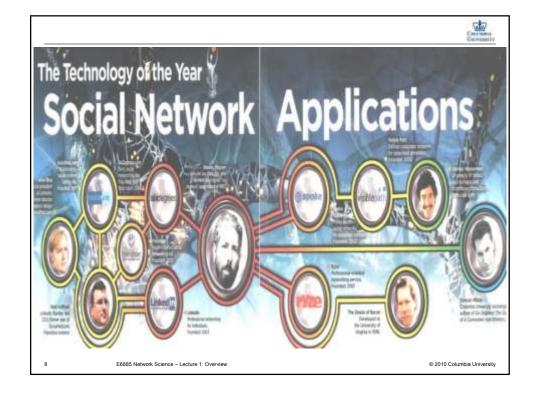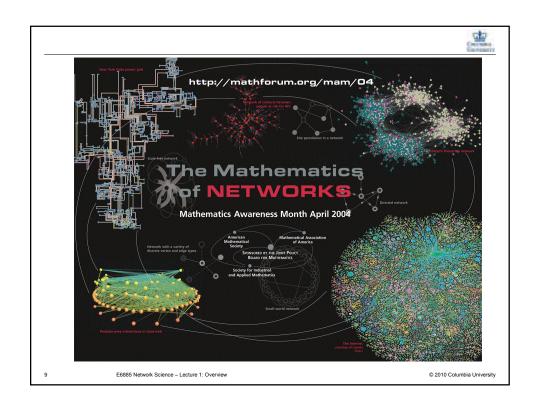
The first series of maps of the new human geography were shown by Dr. Jacob L. Moreno of New York, consulting psychiatrist of the National Committee of Prisons and Prison Labor and director of research, New York State Training School for Girls, Hudson, N. Y. The maps represent studies of the forces of attraction and repulsion of individuals within a group toward one another and toward the group, as well as the attitude of the group as a whole toward its individual members, and of one group toward another group.
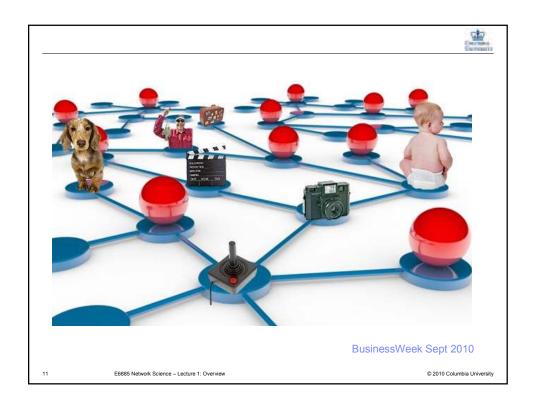
Emotions are represented on these psychological maps by various colored lines. Red stands for liking, black for disliking. If individual A likes B a red line with an arrow points from A to B. If B reciprocates a similar red line points from him. If he dislikes A this is indicated by a black line with an arrow pointing toward A. If B is merely indifferent the feeling is shown by a blue line.

Group of 500 Girls Studied.

The Technology of the Year
Social Network Applications

4

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

---
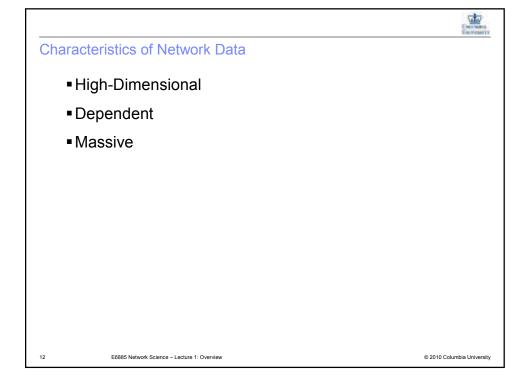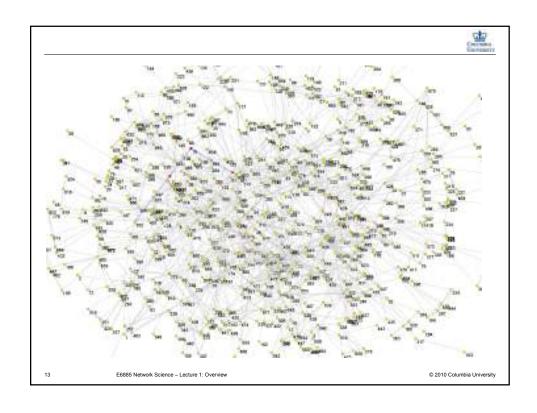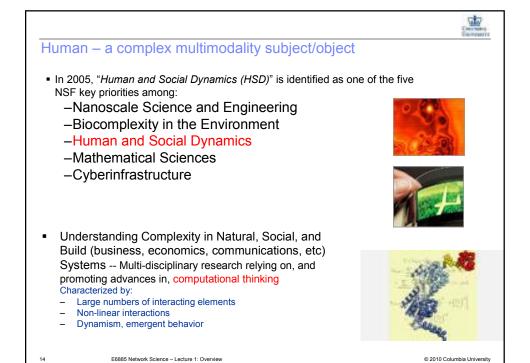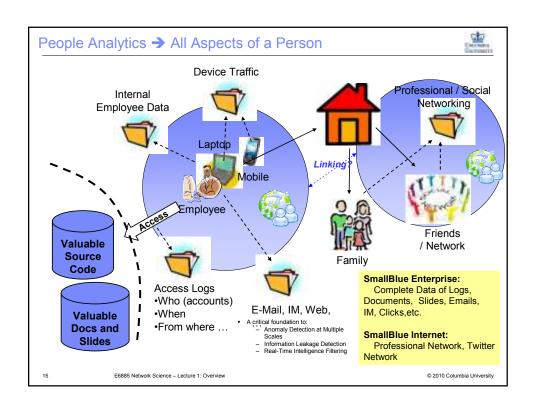
Booming of Online Social Networks

- Chomping at the bit for a **Facebook** I.P.O., investors are buying common stock as high as $76 a share, giving the company a value of as much as $33,700,000,000 based on secondary market transactions, The Financial Times – August 2010.

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

5

BusinessWeek Sept 2010

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

---

Characteristics of Network Data

- High-Dimensional
- Dependent
- Massive

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

---

Human – a complex multimodality subject/object

- In 2005, "*Human and Social Dynamics (HSD)*" is identified as one of the five NSF key priorities among:

  – Nanoscale Science and Engineering
  – Biocomplexity in the Environment
  – Human and Social Dynamics
  – Mathematical Sciences
  – Cyberinfrastructure

- Understanding Complexity in Natural, Social, and Build (business, economics, communications, etc) Systems -- Multi-disciplinary research relying on, and promoting advances in, computational thinking
  Characterized by:
  - Large numbers of interacting elements
  - Non-linear interactions
  - Dynamism, emergent behavior

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

People Analytics ➜ All Aspects of a Person

Device Traffic

Internal Employee Data

Professional / Social Networking

Laptop

Mobile

Linking?

Employee

Access

Valuable Source Code

Valuable Docs and Slides

Access Logs
•Who (accounts)
•When
•From where …

E-Mail, IM, Web,
• A critical foundation to:
  – Anomaly Detection at Multiple Scales
  – Information Leakage Detection
  – Real-Time Intelligence Filtering

Family

Friends / Network

SmallBlue Enterprise:
  Complete Data of Logs, Documents, Slides, Emails, IM, Clicks,etc.

SmallBlue Internet:
  Professional Network, Twitter Network

15   E6885 Network Science – Lecture 1: Overview   © 2010 Columbia University

---

Cognitive Networks

Cognitive Network

- Cognitive Network:
  – 30,000 nodes of dynamic brain MRI functional networks
16   E6885 Network Science – Lecture 1: Overview   © 2010 Columbia University

Composite Social-Cognitive-Info Networks

3G

3G

3G

Status Monitoring,

Visualization, Personalized Information Recommendation, Routing, etc.

EEG / Audio signal detection

GPS / location detection,

Information Display

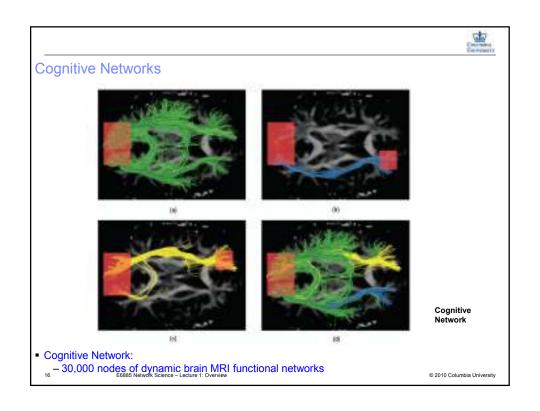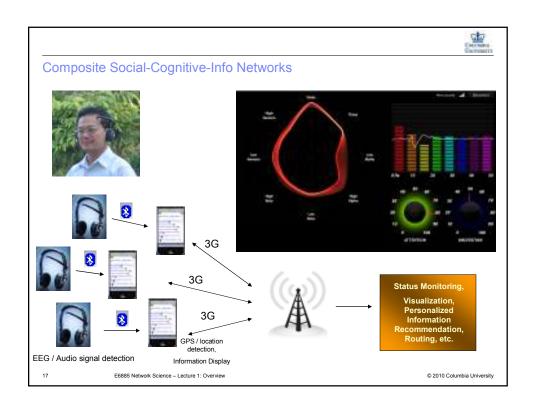E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

---

The Emergence of Network Science

- Science <=> Observable systematic empirical data

- Facility of large-scale data collection, storage and management.

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

The Emergence of Network Science

- Science <=> Observable systematic empirical data

- Facility of large-scale data collection, storage and management.

    - *Statistical Methodologies to combine behavior understanding, link analysis, multi-variant modeling, machine learning, graph theory, and non-parametric statistics for complex network analysis*

> Contributions made by Physicists and Computer Scientists, have greatly expanded the discipline over the past 15 years.

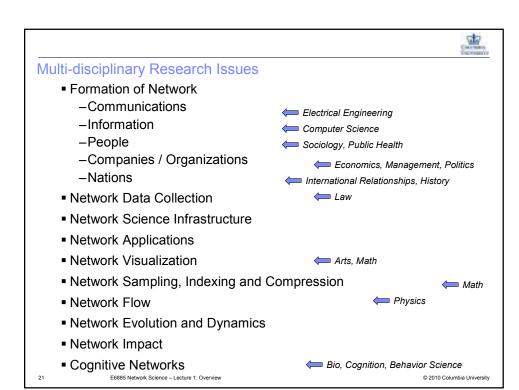## Multi-disciplinary Research Issues

- Formation of Network
    - Communications
    - Information
    - People
    - Companies / Organizations
    - Nations
- Network Data Collection
- Network Science Infrastructure
- Network Applications
- Network Visualization
- Network Sampling, Indexing and Compression
- Network Flow
- Network Evolution and Dynamics
- Network Impact
- Cognitive Networks

Multi-disciplinary Research Issues

- Formation of Network
    - Communications   ⬅ *Electrical Engineering*
    - Information   ⬅ *Computer Science*
    - People   ⬅ *Sociology, Public Health*
    - Companies / Organizations   ⬅ *Economics, Management, Politics*
    - Nations   ⬅ *International Relationships, History*
- Network Data Collection   ⬅ *Law*
- Network Science Infrastructure
- Network Applications
- Network Visualization   ⬅ *Arts, Math*
- Network Sampling, Indexing and Compression   ⬅ *Math*
- Network Flow   ⬅ *Physics*
- Network Evolution and Dynamics
- Network Impact
- Cognitive Networks   ⬅ *Bio, Cognition, Behavior Science*
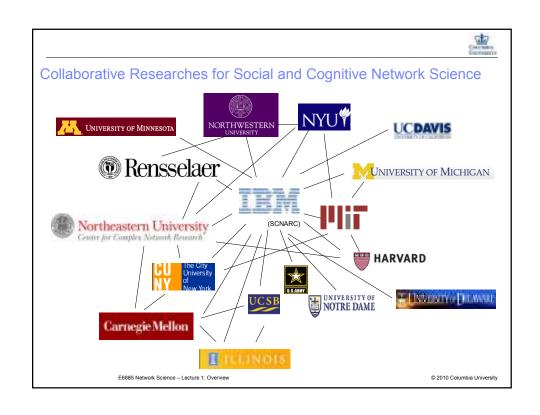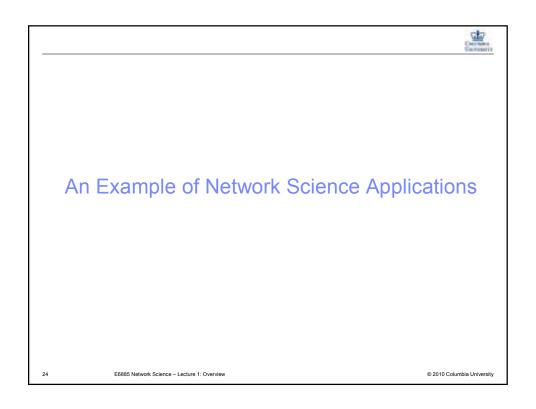

Example: Network Science Consortium (2009 – 2014 or 2019)

- U.S. Army is investing $172 million in 10 years to fund 4 *Network Science* academic research centers:
    - Communication Network Academic Research Center (CNARC)
    - Social and Cognitive Network Academic Research Center (SCNARC)
    - Information Network Academic Research Center (INARC)
    - Interdisciplinary Research Center (IRC)

    ~ 100 Professor/Researcher Principle Investigators + 250 RAs, Postdocs

- Objectives:
    - Improve Decision Making
    - Develop measures, metrics and models that describe and predict human-network interaction and exchange within & across network layers
    - Develop and validate theory of human-system interaction in network-centric environments
    - Explore techniques for dynamic, flexible, adaptive, and adaptable interaction

> 1. Charactering and Measuring Networks
>
> 2. Understanding Networks for Analysis
>
> 3. Controlling and Managing Networks
>
> 4. Using Networks

Collaborative Researches for Social and Cognitive Network Science

An Example of Network Science Applications

Unlocking the expertise that resides among my friends, or my friends' friends…

- "If I know who among our 400,000 IBM colleagues knows the answer of this critical question customer is asking, a million dollar deal signing may be only one phone call away…" *-- an IBM CHQ Sr. VP in 2003, showing how desperately IBM needs a novel expertise location tool…*

Gartner 2003 – Personal Network preferred source for information

Personal Network   Preferred / primary mode

?

Practitioner with task in project / delivery environment

Forces:
- Time Constrained
- Delivery activity focus
- What gets measured gets done
- Expedience
- Perceived value (return on time investment)

High reliance on:
- Personal networks
- Hard-drive materials
- What has worked for them previously (personal experience)

| client | client | W3 Stub | W3 Stub | W3 Stub / client | W3 Stub / Client | W3 Stub / client | W3 Stub | W3 Stub / client |
| Project Tools | Project Repositories | Knowledge View | PSN | Methods | Education | Communities | Other w3 content | Collaboration |

13

Where does knowledge reside?

Emails, Instant Messages, Calendars, Databases, Wikis, Blogs,
Microblogging..

ExpertiseNet: Modeling Person's Expertise Evolution as Dynamic Graphs (Song, Tseng, Lin and Sun, User Modeling 2005)

Data → Text Analysis / Link Analysis → Expertise Extraction / Relationship Representation / Evolutionary Representation → ExpertiseNet → Expertise Mining / Expertise Matching / Pattern Discovery

Build graphical representations of a researcher's expertise based on her/his publications and citation networks

---

SmallBlue asset: Large-Scale People Modeling and Social Network Analysis

Emails / Chats / Meetings / Web Page Clicks / Server Logs

Social sensors / Click streams capturer / Feed subscription / Database access

Network analysis / People analysis / Content analysis

Applications

Live Data, Live Production System

20,000,000 emails & SameTime messages

1,000,000 Learning click data

14,000,000 KnowledgeView, SalesOne, …, access data

1,000,000 Lotus Connections (blogs, flie sharing, bookmark) data

200,000 people's consulting financial databases

400,000 organization/demographic data

100,000 intranet w3 searches per day

15

SmallBlue unlocks the Power of Business Networks and Protects Privacy

Expertise: Search for people who know "xyz" in my networks..

Net: See how experts or community connect

Ego: Show my personal network evolution and social capital

Social Network Analysis & Visualization, Expertise Mining, and Multi-Channel Human Network/Behavior Analysis

Reach: helps me understand this person, my formal and informal paths to Reach him..

Whisper: Social Network enabled personalized live recommender..

Synergy: Personalized Search

Productivity: Social Network Analysis Service helps company understand how to enhance productivity.

Global Privacy Laws

European Union
• European Data Protection Directive (1995)

Russia
• Federal law on Pers Data (January 2007)

Canada
• PIPEDA (2001 - 2004)

APEC
• Guidelines (2004)

APEC
• Guidelines (2004)

Australia
• Privacy Amendment Act (2001)

New Zealand
• Privacy Act (1993)

U.S. – Sectoral
• Children's Privacy; COPPA (1999)
• Financial Sector GLB (2001)
• Health Sector; HIPAA (2002)
• California Privacy; (2005)

Dubai
• Data Protection Law (January 2007)

Taiwan
• Computer - Processed PD Protection Law (1995)

■ *Existing Private Sector Privacy Laws*

Chile
• Protection of Private Life Law (1999)

South Korea
• Info & Comm Network Util. & Info Protection Law (2000)

■ *Emerging Private Sector Privacy Laws*

Argentina
• Protection of PD Law (2000)

Japan
• Personal Data Protection Act (2005)

16

SmallBlue Applications (I): Find knowledgeable colleagues

- E.g.: Search for the most knowledgeable colleagues within my 3-degree network for who knows 'healthcare'. (or within a country, a division, a job role, or any group/community)

My shortest path to Susan

As a user, you can only see their public information. Private info is used internally to rank expertise but private data can never be exposed.

Click a name to see their profile (SmallBlue Reach)

33   E6885 Network Science – Lecture 1: Overview   © 2010 Columbia University

SmallBlue Applications (II): Reach – social dashboards

- Is Tom a right person to me?

His official job role, title, contact info

His self-described expertise

His public communities

The public interest groups he is in

His blogs, forum, postings..

My various paths to Tom. SmallBlue can show the paths to any colleagues up to 6-degree away

34   E6885 Network Science – Lecture 1: Overview   © 2010 Columbia University

17

SmallBlue Applications (II): Reach (cont'd)

- **Is Tom a right person to me?**

Formal hierarchy path, if I really want to go through this 13-degree path…

The network of people we know in-between us.

35   E6885 Network Science – Lecture 1: Overview   © 2010 Columbia University

SmallBlue Applications (III): Net – corporate social network analysis

- **How are company's top healthcare experts link with each other? Who are the key bridges? Who have the most connections? How do these experts cluster?**

Independent experts on healthcare

A cluster of healthcare experts

Top-N experts on healthcare

Highlight experts based on my social proximity, the number of experts she connects, or the 'social bridges' importance

36   E6885 Network Science – Lecture 1: Overview   © 2010 Columbia University

18

SmallBlue Applications (III): Net (cont'd)

Healthcare experts in the world

Connections between different divisions

Healthcare experts in the U.S.

Key social bridges

SmallBlue Applications (IV): Ego – personal social network capital management [an application only visible to the user himself]

- What is a friend's social capital to me?

It can also show the evolution of my social network..

How many people in my personal networks?

SmallBlue finds dynamic existing social networks of every employee That makes it possible to find the shortest path to any colleague..

What types of unique colleagues my friend Chris can help me connect to?

My personal social network automatically found by SmallBlue with social distance

Personalized Content Recommendation and Search

- What your friends know become what you know.. Your friends are your window to the world – *Confucius ~600 B.C.*
- Utilizing the unique large-scale weighted social network inferred by SmallBlue, personalized ranking becomes possible.
- Fusion of Recommenders: Social Filtering, Collaborative Filtering, Latent Semantic Filtering, Popularity & Freshness Filtering, etc.
- Deploy on IBM KnowledgeView, IBM Learning, and IBM TAP

SmallBlue Whisper

Ching-Yung Lin
IBM Research
Complex and Social Network Anal

46 IBMers are 1 degree away from you.
1278 IBMers are 2 degrees away from you.
14749 IBMers are 3 degrees away from you.

- **SmallBlue Whisper and Synergy usages:**
  - **On Nov. 19, 2008: 11,108**
  - **On Nov. 20, 2008: 9,567**

Current search

documents → tagging → ranking → index → re-ranking → re-index → Improved search results / Interest / social network based content recommendations

Metadata
Formal taxonomy / social tags

understanding of the user

SmallBlue Synergy

Link analysis

Get understanding of the users interests based on:
- Keywords from SmallBlue communication analysis
Recommend content based on data

Synergy – Personalized Content Search

SmallBlue (Atlas) Mobile

- SmallBlue Applications on Mobile Phones

Android

BlackBerry

Nokia

Show Expertise of 'SNA' inside:
(1) IBM
(2) My 2-degree network
(3) Research division
(4) Global Business Services
(5) Any group – e.g., Distinguished Engineers

Recommend Contents from Friends within 3-degrees

SmallBlue Find Widget in Mobile

SmallBlue Whisper Widget in Mobile

20

Several Technical Challenges

- Large-Scale Expertise ranking – Google PageRank version of People (This is not something in open sources such as Lucene)

- Real-Time Dynamic Large-Scale Social Path Graph Indexing

- Distributed Privacy-Preserving Sensor System **– what personal data shall be executed in personal control; what personal data can be sent to and processed by server..**

- Requirements for expertise/knowledge search:
  – Dirty word detector
  – Name detector
  – Sensitive query detector (dollars, confidential words, numbers, etc.)
  – Multi-language processing (Languageware components)
  – Multi-platform (Lotus Domino Server, Microsoft Exchange Server, Google/Yahoo/MSN imap server)
  – Multi-search requirements (Job roles, geography, community, any BlueGroups)

- Scalability **– 7 nodes for production, 4 nodes for staging, for IBM deployment. Flexible design for different company sizes.**

- Stability **– transferred to Lotus Software Developers in 2007. Compliant with IBM Software development standards. Full documentations (30 architecture documents). Production Services since March 2008 – 18 months already.** 713,000 usages in the last 11 months.

- Accessibility **-- For Blind and Disabled Users**

- System configurability and management **– Web-based admin reports, Websphere configuration components..**

Network = Graph + Attributes

Graph Definitions and Concepts

- A graph:

$$G = (V, E)$$

- V = Vertices or Nodes
- E = Edges or Links

- The number of vertices: "Order" $N_v = |V|$

- The number of edges: "Size" $N_e = |E|$
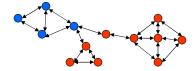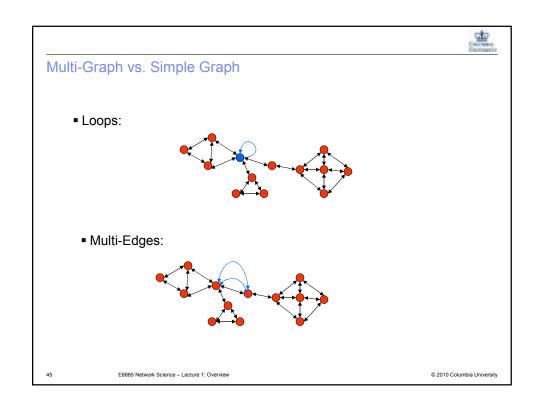
Subgraph
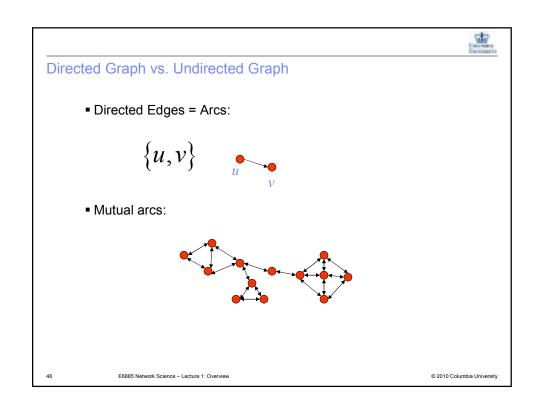
- A graph H is a subgraph of another graph G, if:

$$V_H \subseteq V_G \quad \text{and} \quad E_H \subseteq E_G$$

Multi-Graph vs. Simple Graph

- Loops:

- Multi-Edges:

Directed Graph vs. Undirected Graph

- Directed Edges = Arcs:

$$\{u, v\}$$

u v

- Mutual arcs:
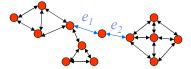
Adjacency

- *u* and *v* are adjacent if joined by an edge in E:

- Two edges are adjacent if joined by a common endpoint in V:
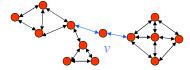
Incident and Degree

- A vertex $v \in V$ is incident on an edge $e \in E$ if v is an endpoint of e.

- The degree of a vertex v, say d_v, is defined as the number of edges incident on v.

$d_v = 2$

24
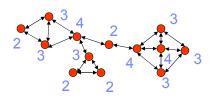
Degree Sequence

- The degree sequence of a graph G is the sequence formed by arranging the vertex degrees d_v in non-decreasing order.

$$\{2,2,2,2,3,3,3,3,3,3,4,4,4\}$$

- The sum of the elements degree sequence equals to twice the number of edges in the graph (i.e. twice the size of the graph).

In-degrees and out-degrees

- For Directed graphs:

In-degree = 8   Out-degree = 8

25
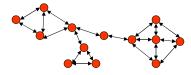
Walk

- A walk on a graph G, from v_0 to v_l, is an alternating sequence:

$$\{v_0, e_1, v_1, e_2, ..., v_{l-1}, e_l, v_l\}$$

- The length of this walk is l.
- A walk may be:
  - Trail --- no repeated edges
  - Path --- trails without repeated vertices.
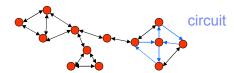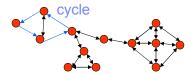
E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

---

Circuit, Cycle, and Acyclic

- Circuit: A trail for which the beginning and ending vertices are the same.

circuit

- Cycle: a walk of length at least three, the beginning node = ending node, all other nodes are distinct

cycle
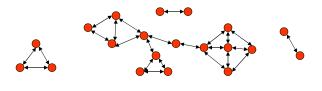
- Acycle: graph contains no cycle

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

Reachable, Connected, Component

- Reachable: A vertex v in a graph G is said to be reachable from another vertex u if there exists a walk from u to v.

- Connected: A graph is said to be connected if every vertex is reachable from every other.

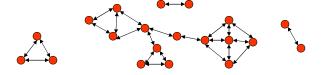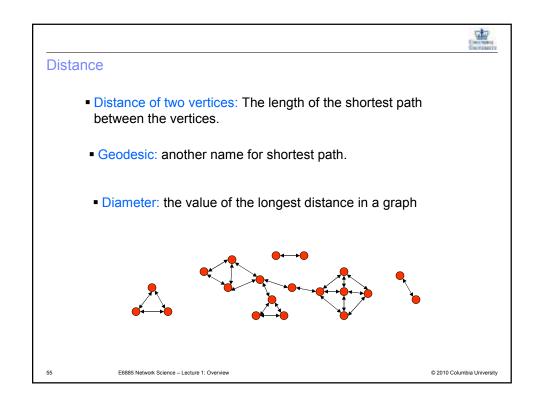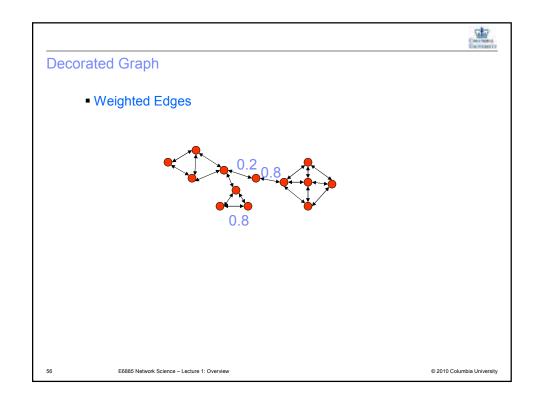- Component: A component of a graph is a maximally connected subgraph.
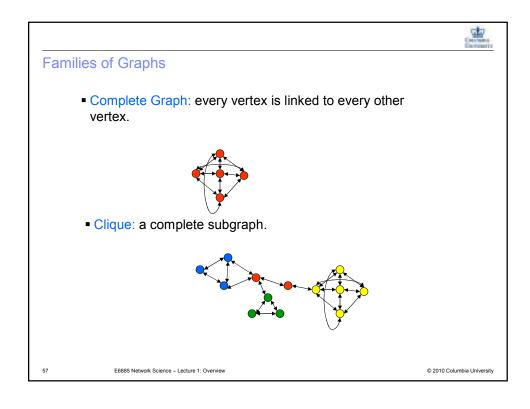
Connection in a digraph

- Weakly connected: If its underlying graph is connected after stripping away the direction.

- Strongly connected: every vertex is reachable from every other vertex by a directed walk.
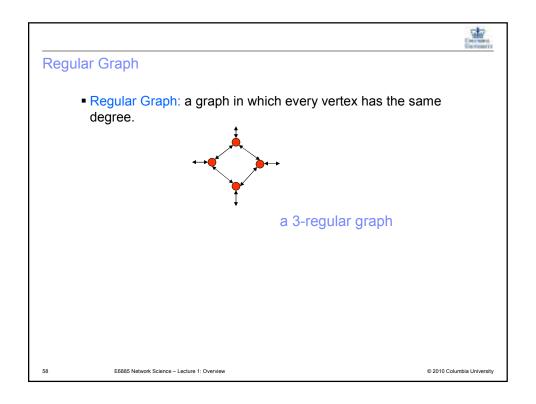
Distance

- Distance of two vertices: The length of the shortest path between the vertices.

- Geodesic: another name for shortest path.

- Diameter: the value of the longest distance in a graph

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

---

Decorated Graph

- Weighted Edges

0.2 0.8

0.8

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

Families of Graphs

- Complete Graph: every vertex is linked to every other vertex.

- Clique: a complete subgraph.

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

Regular Graph

- Regular Graph: a graph in which every vertex has the same degree.

a 3-regular graph

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

Tree and Forest

- **Tree:** a connected graph with no cycle.

- **Forest:** a disjoint union of trees is called a forest.
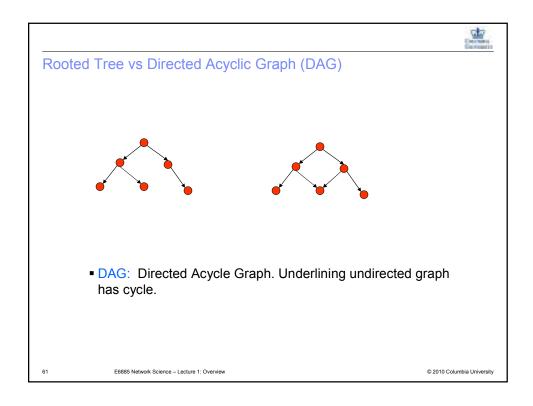
Labels in a directed tree
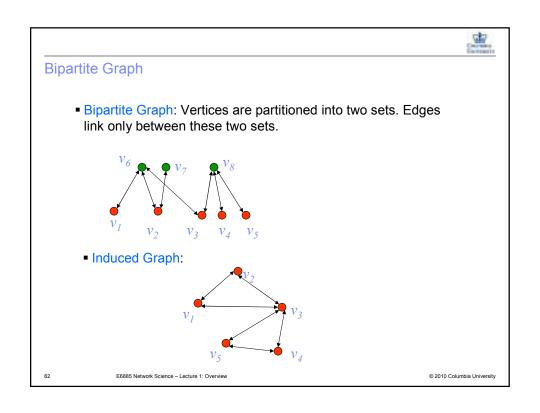
- **Root**
- **Ancestor**
- **Descendant**
- **Parent**
- **Children**
- **Leaf:** a vertex without children

Rooted Tree vs Directed Acyclic Graph (DAG)

- DAG: Directed Acycle Graph. Underlining undirected graph has cycle.

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

---

Bipartite Graph

- Bipartite Graph: Vertices are partitioned into two sets. Edges link only between these two sets.

- Induced Graph:

E6885 Network Science – Lecture 1: Overview © 2010 Columbia University

Graphs and Matrix Algebra

- The fundamental connectivity of a graph *G* may be captured in an $N_v \times N_v$ binary symmetric matrix *A* with entries:

$$A_{ij} = \begin{cases} 1, & if \{i, j\} \in E \\ 0, & otherwise \end{cases}$$

A is called the Adjacency Matrix of *G*

Questions?