Lecture 14: Search and Navigation on Social and other Networks: Affiliation Hierarchies

Christopher L. Magee

April 4, 2006

Lecture 14 Outline

- Introductory Remarks
- Search and Navigation, search (briefly)
- Navigation
  - Milgram's experiment and critiques
  - Kleinberg's first model
  - The influence of structure and Kleinberg's second model (and the Watts, Dodds and Newman model)
- Modeling Overview
  - Search/navigation as case study of model evolution
  - Modeling limitations and benefits
  - The fundamental modeling tradeoff
  - Systems and applications of most interest

Search and Navigation

- Search
  - "To look over carefully in order to find something, to explore", "to make an effort to find something" seek, hunt, quest.

- Navigate
  - "To plan, record and control the position of.." "to follow a planned Course" or "*to make one's way*"

- Use of maps is not explicitly mentioned but their usage relative to navigation seems more normal than for search.

Search and Navigation

- Search
  - "To look over carefully in order to find something, to explore", "to make an effort to find something" seek, hunt, quest.
  - Network literature: **"to find the node containing information that is desired"**

- Navigate
  - "To plan, record and control the position of.." "to follow a planned Course" or "*to make one's way*"
  - Network literature: "**to get from one to another specific node by a(the) short(est) path using *only local information*"**

Network Search I

- Exhaustive WWW Search
    - Catalog (while "crawling") the network and create a map (local index) of the **entire network**
    - Use information in nodes to select relevant web pages
    - Rank nodes for significance using the **link** information
        - Eigenvector Centrality (Brin and Page)
        - Each node has a weight x_i that is defined to be proportional to the weights of all nodes that point to i
        - And

            $$x_i = \lambda^{-1} \sum_j A_{ij} x_j$$

        - And then $Ax = \lambda x$
        - Thus the weights are an eigenvector of the adjacency matrix (A) with eigenvalue λ

        Kleinberg has considered a more sophisticated version with weights for *pointing* hubs and *receiving* hubs

Network Search II

- Guided Search- databases or the web for unmapped elements
  - Web "spiders"
  - Queries passed (if not answered) to highest k node of neighbors. If some nearest neighbor information is also stored this is a valuable approach for peer-to-peer systems
  - This last approach takes advantage of structure with knowledge of high k nodes being present

Network Navigation: Milgram's experiment

- The unpublished (but widely circulated) paper of Kochen and Pool using simple random graph models indicated the possibility of short paths through social networks. This instigated the famous social scientist Stanley Milgram to try an experiment

- "route" letter to person XXX who is a stockbroker living in Sharon, MA who works in Boston.

- The letters can ***only*** be sent to someone who the recipient knows on a ***first name*** basis but in a way to get "closer" to person XXX. Participants also were asked to record and send along the routing information

Social networks: Milgram's experiment

Figure removed for copyright reasons.
Source: Milgram, *Psych Today* **2**, 60 (1967).

Milgram's experiment

- "route" letter to person XXX who is a stockbroker living in Sharon, MA who works in Boston.

- The letters can **only** be sent to someone who the recipient knows on a **first name** basis but in a way to get closer" to person XXX.

- Some guesses that it would take hundreds of steps were refuted by results that "showed" it took (**actually can take**) much less

Six-degrees of separation

Distance

Figure removed for copyright reasons.
Source: Milgram, *Psych Today* **2**, 60 (1967).

Milgram's experiment II

- "route" letter to person XXX who is a stockbroker living in Sharon, MA who works in Boston.
- The letters can **only** be sent to someone who the recipient knows on a ***first name*** basis but in a way to get closer" to person XXX.

- Some guesses that it would take hundreds of steps were refuted by results that "showed" it took (actually can take) much less
- A play was written and coined the phrase "*six degrees of separation*" as its Title and Milgram's result became something "everyone knows".
  - ***Everyone*** is separated by only six removes from ***everyone*** else on the planet!
- But What did Milgram really show?

What did Milgram really show?

- Of 300 letters in original experiment, only 96 (random Nebraska) sampled tested the "*everyone*" part of what "*everyone*" knows

- Only 18 of these were ***ever returned*** (the preceding graph contained very "non-random" Nebraska letters)

- Other trials that were random and tested the everyone basis had even ***smaller return rates*** than Milgram's initial experiment

- Issues
  - Is *everyone* really 6 or less steps from *everyone*?
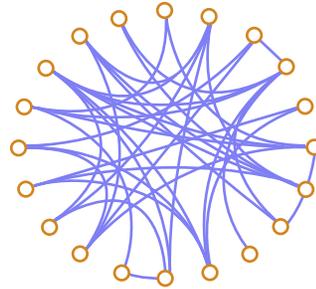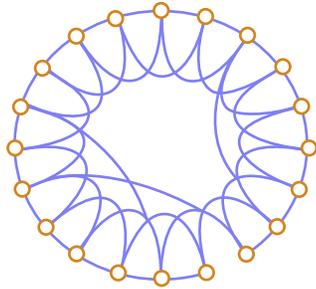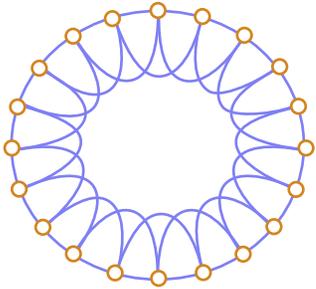
Small-world networks

Regular	Small-world	Random

Increasing Randomness

p = 0 → p = 1

N = 1000

$C(p) / C(0)$

$L(p) / L(0)$

- Large clustering coeff.
- Short typical path

Watts & Strogatz,

Nature **393**, 440 (1998)

p

Massachusetts Institute of Technology

Ubiquity of small-world networks

$C(p) / C(0)$

N-dependence?

$L(p) / L(0)$

p
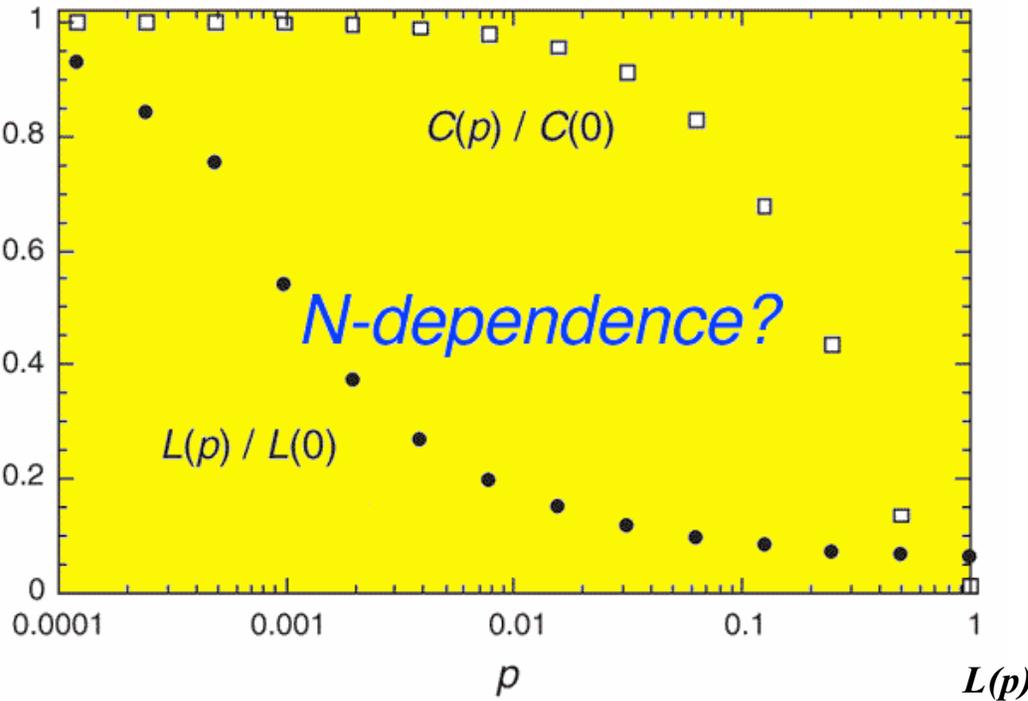
Figure by MIT OCW. After Watts & Strogatz, 1998.

$$L(p, N) \sim N_* \, \mathrm{F}\!\left(\frac{N}{N_*} \right)$$

Figure by MIT OCW. After Barrat & Weigt, 2000.

N=100

N=20000

$$N_* \approx \frac{1}{p}$$

$L(p)$

p

Bertelemy and Amaral, Phys Rev Lett **83**, 3180 (1999)

Newman & Watts, *Phys Lett A* **263**, 341 (1999)
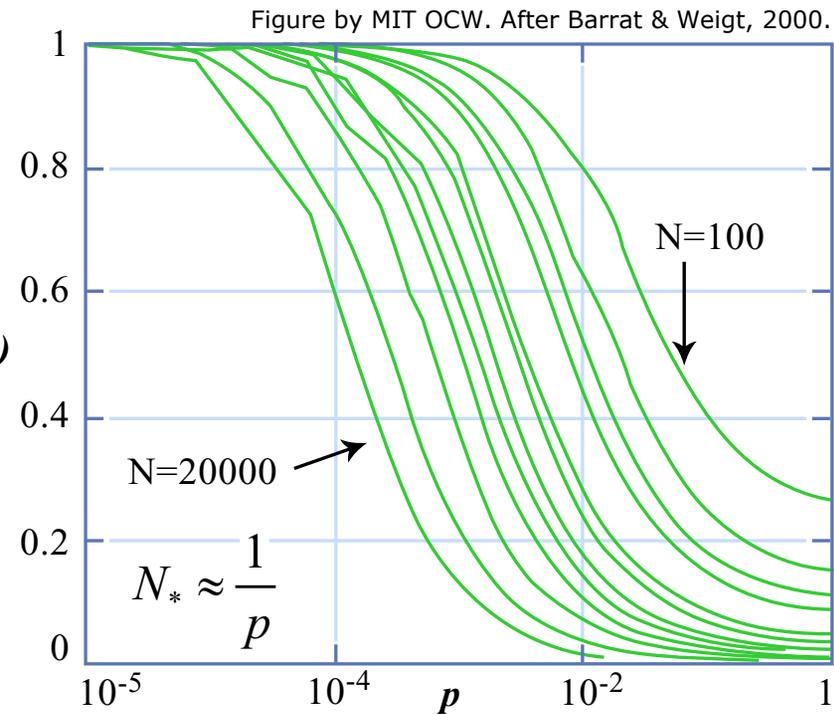
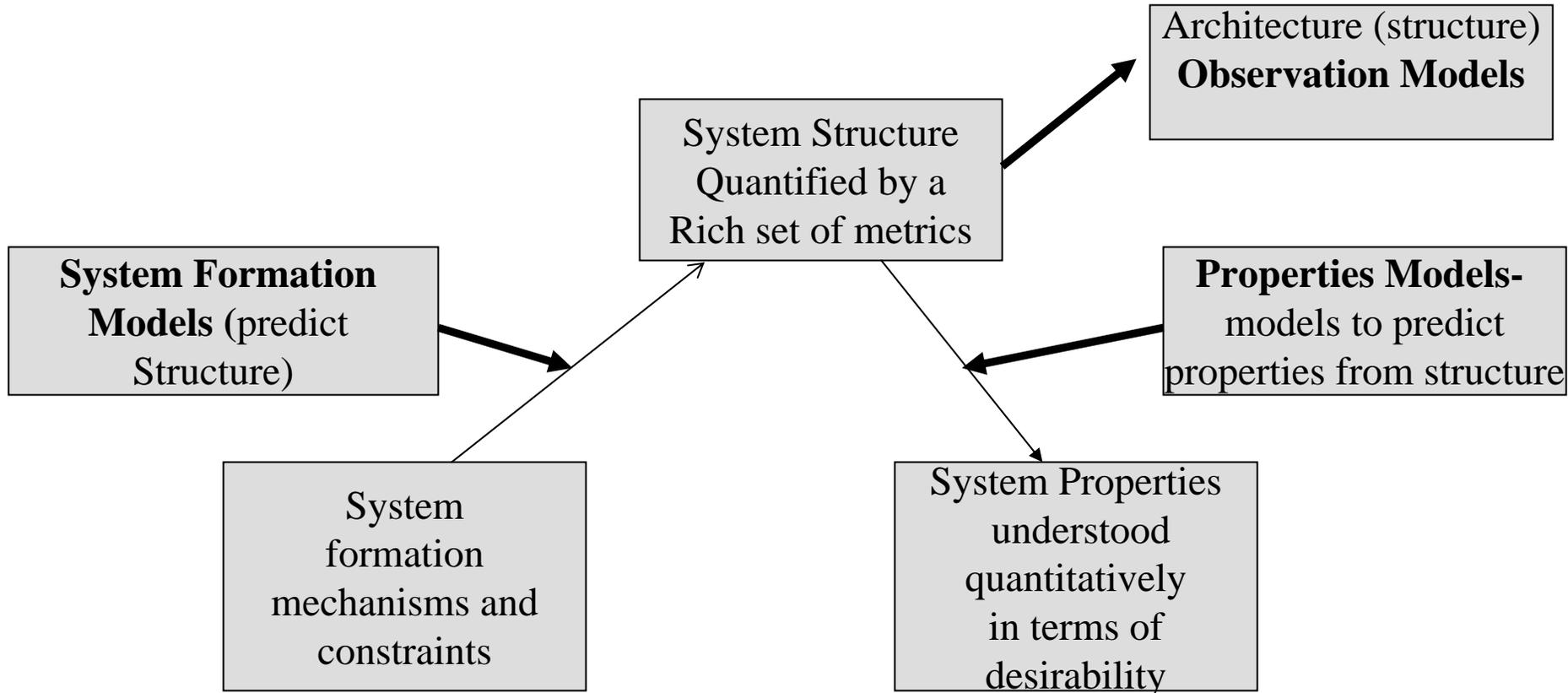Barrat & Weigt, *Eur Phys J B* **13**, 547 (2000)

Potential short paths

- There are almost surely *relatively* short paths between **any** two individuals

- The path length is *apparently* about that calculated for random networks:

$$l \cong \frac{\ln n}{\ln <k>}$$

- For n representative of the whole world, this would give path lengths as large as 10-20. Even though 10 degrees of separation does not sound as impressive, it is still small.
- As a model, the Small World Model is obviously primitive as a "Systems Formation" Model. For this phenomena/purpose (explaining Milgram's experiment), is this its most serious shortfall?

Schematic of Engineering System Model Types within a Framework

Architecture (structure)
Observation Models

System Structure Quantified by a Rich set of metrics

System Formation Models (predict Structure)

Properties Models- models to predict properties from structure

System formation mechanisms and constraints

System Properties understood quantitatively in terms of desirability

What did Milgram really show?

- Of 300 letters in original experiment, only 96 (random Nebraska) sampled tested the "*everyone*" part of what "*everyone*" knows
- Only 18 of these were *ever returned* (the preceding graph contained very "non-random" Nebraska letters)
- Other trials that were random and tested the everyone basis had even *smaller return rates* than Milgram's initial experiment

- Issues
  - How does *anyone* route such a request?
  - Is *everyone* really 6 or less steps from *everyone*?
  - Apathy vs. possibility –even harder now that we all toss/delete junk mail and return rates near 1% are apparently now happening with Watts's e-mail small world experiment.

Kleinberg's initial model

- Most important insight
  - Milgram's experiment did not only show that short paths exist but more importantly *that people can* (at least sometimes and in some circumstances) *find them*.
- Model assumptions
  - Small World (with shortcuts added onto a **lattice** of connections) –**not randomly** but with a probability that depends on distance,

Massachusetts Institute of Technology

"Navigation in Small Worlds: It is easier to find short chains in some networks than others"
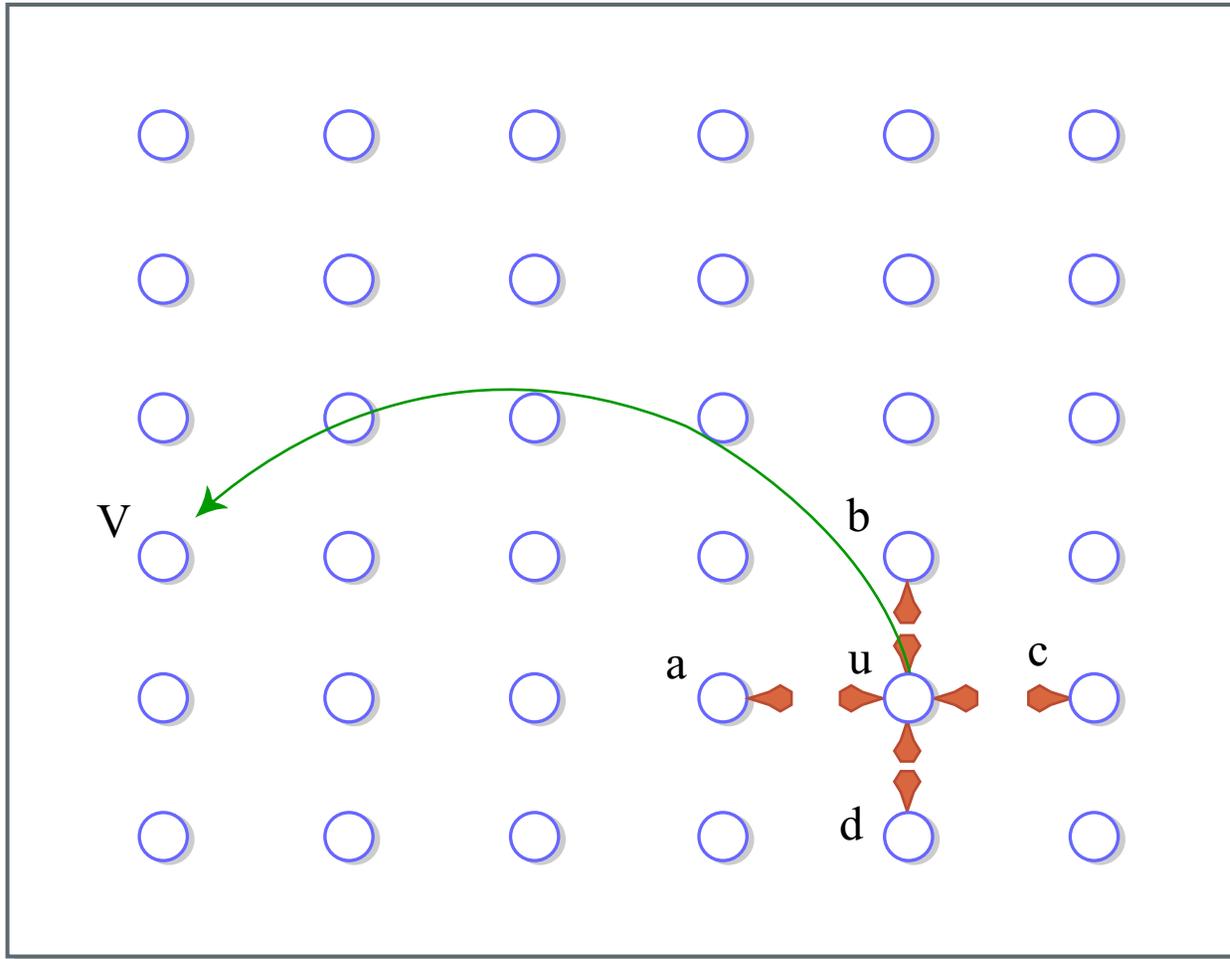

Figure by MIT OCW.

Massachusetts Institute of Technology

Kleinberg's initial model

- Most important insight
  - Milgram's experiment did not only show that short paths exist but more importantly ***that people can*** (at least sometimes and in some circumstances) ***find them***.
- Model assumptions
  - Small World (with shortcuts added onto a **lattice** of connections) –**not randomly** but with a probability that depends on distance,

  - $$p_s \propto r^{-\alpha} \quad \text{steps to find} \quad S_{mean} \geq cn^{\beta}$$

  $$\beta = (2-\alpha)/3 \; \text{for} \; \alpha < 2 \quad \text{and}$$
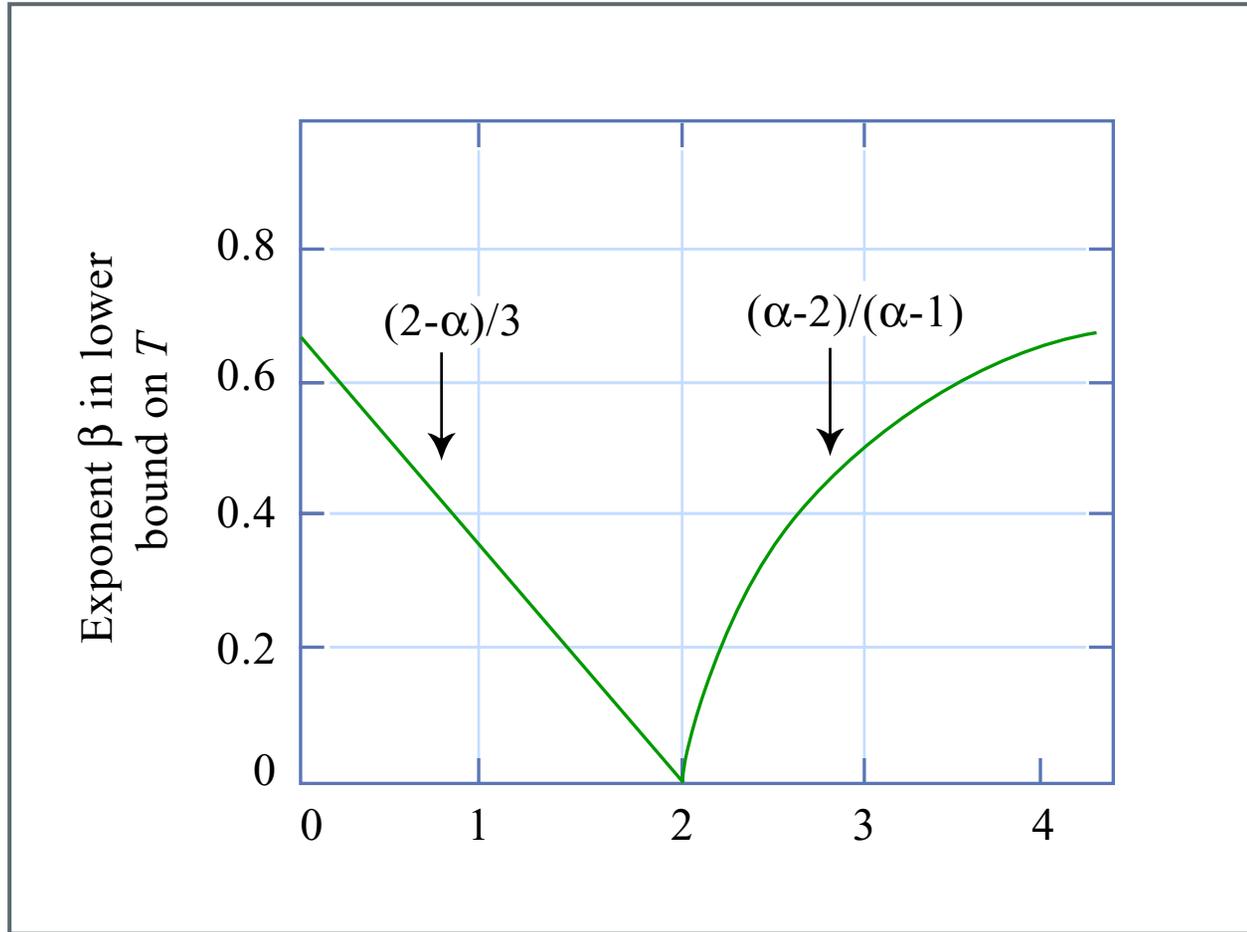 $$\beta = (\alpha-2)/(\alpha-1) \; \text{for} \; \alpha > 2$$

Figure by MIT OCW.

Results of Kleinberg I

- The existence of short paths does not guarantee that they can be found with local information

- It takes network structure of a certain kind ($\alpha = 2$) to be able to do this and to get Milgram's result

- The structure Kleinberg showed worked seems quite artificial but it was a start because it showed that networks can be designed that allow for rapid search with "greedy" algorithms based on local information ("gossip" algorithms)

- Based on this work, even if *everyone* was connected to *everyone*, it is surprising that *anyone* could find the short path.
    - Thus Milgram's famous result is *not explained* by this model

Next Generation structural models for navigation

- Kleinberg and independently Watts, Dodds and Newman have now proposed *a structure* that allows such search and seems consistent with social networks.

- This structure is derived starting from clues from the "Reverse Small World Experiments" which indicate how people actually navigate social networks
  - by looking for common "features" between their targets and their acquaintances

- This structure introduces *hierarchy into the social network* and defines a "social distance".
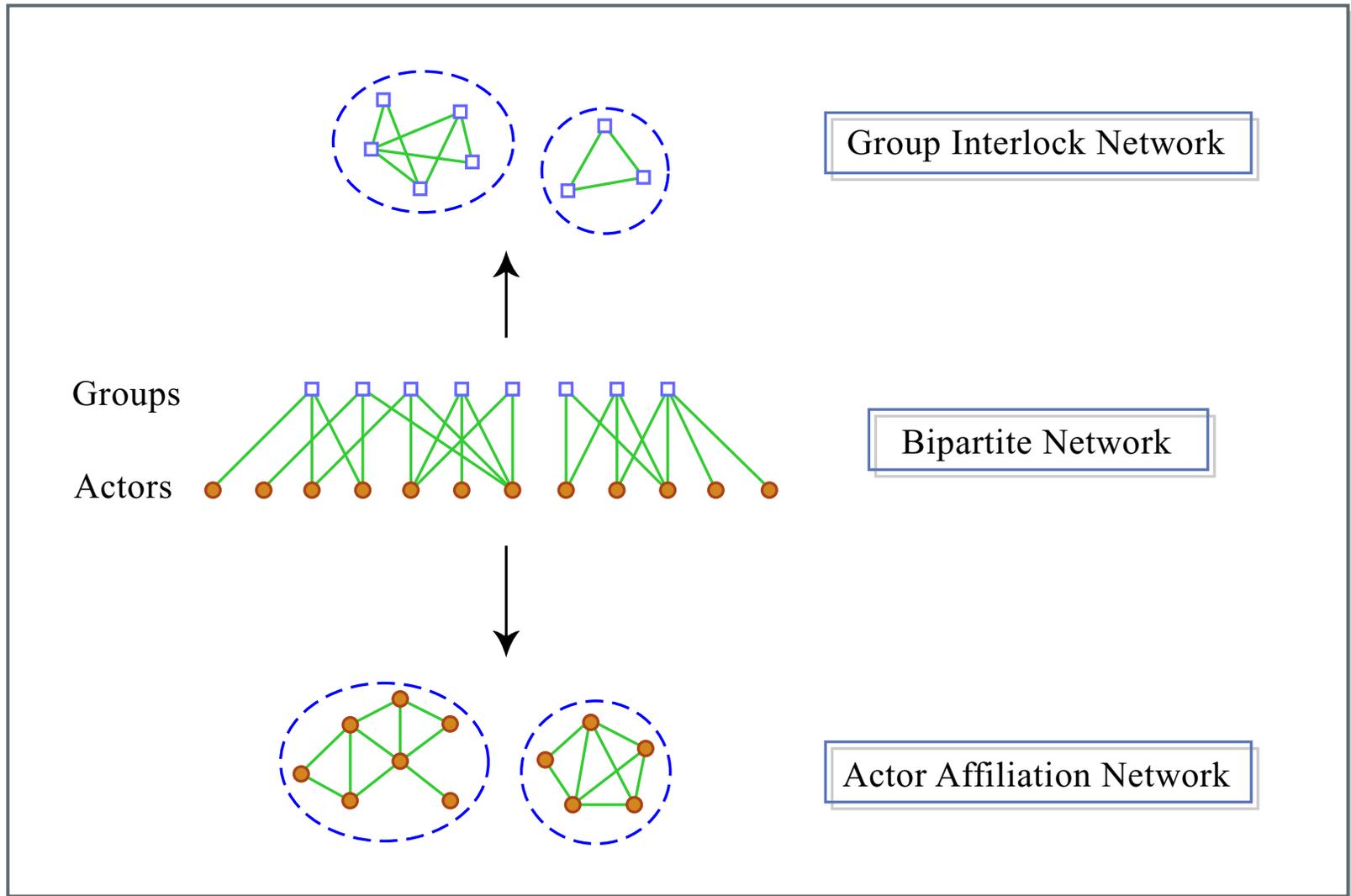
Basic model structure

Group Interlock Network

Groups

Actors

Bipartite Network

Actor Affiliation Network

Figure by MIT OCW. After Watts.
Source: *Six Degrees: The Science of a Connected Age*, Duncan J. Watts, Fig. 4.6, 2003

- 1.   Individuals have ***links and identities***
- ***2.***   Individuals partition the world (identities of others) into a layered hierarchy and ***distance,*** x_{ij} is assumed to be the height of the lowest common parent. The branching ratio, b, and levels, *l* define this abstraction

Figure by MIT OCW.

Massachusetts Institute of Technology

Assumptions in 2ndG Navigation/Search Models II

- 1. Individuals have ***links and identities***

- ***2.*** Individuals partition the world into a layered hierarchy and ***distance***, x_{ij} is assumed to be the height of the lowest common parent. The branching ratio, b, and levels, *l* define this abstraction
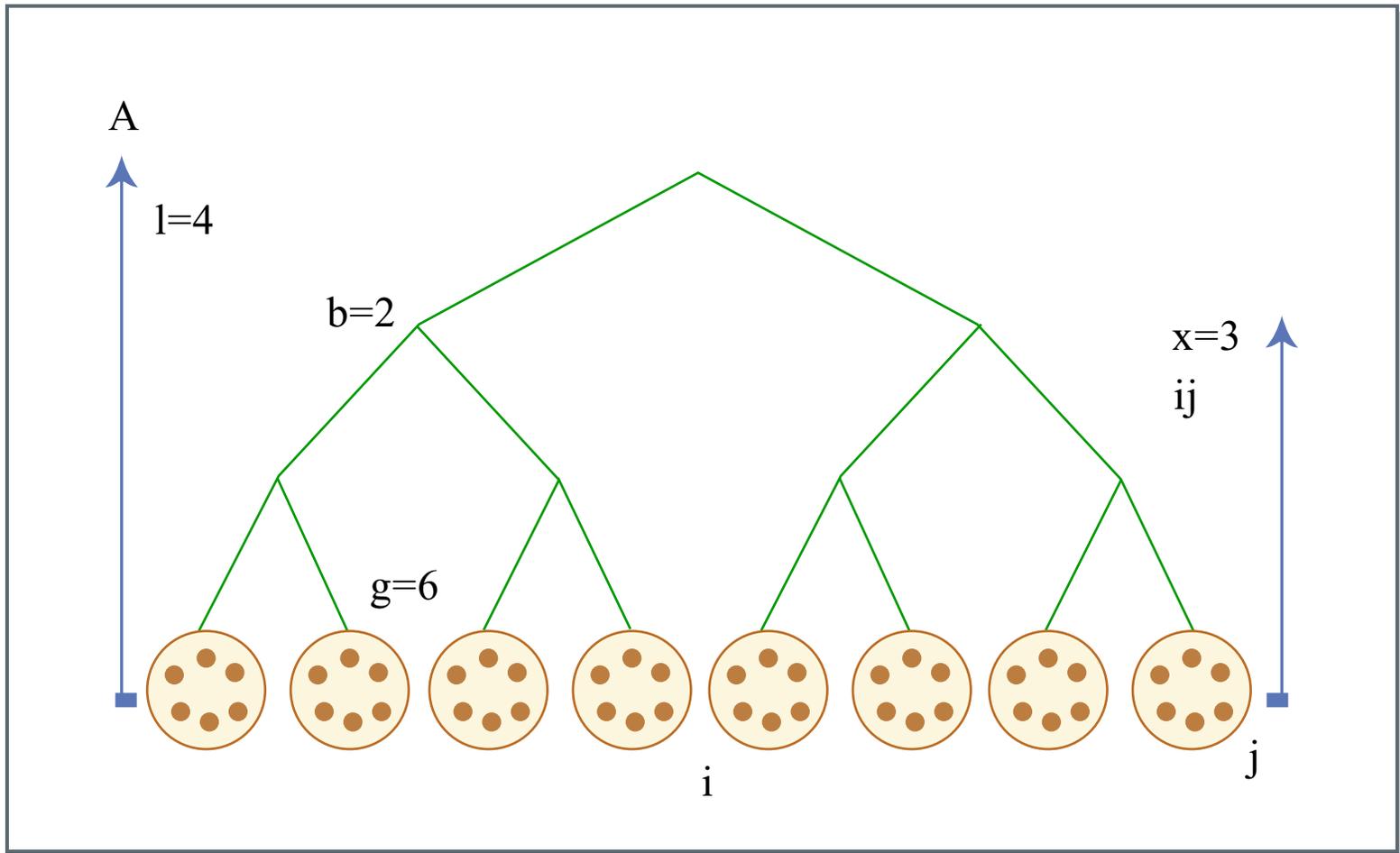
- 3. Group membership signifies not only identity but also is a primary basis for determining social interaction: $p_x = c \exp[-\alpha x]$

- 4. Individuals hierarchically partition the world in more than one way and the model first assumes these distinctions are independent (Kleinberg shows this assumption can be relaxed with qualitatively similar results)

- 5. Individuals construct a measure of "social distance" which is the minimum over all dimensions between the nodes (violates the triangle inequality)

H = 1

Dimension 1
(e.g., geography)

H = 2

Dimension 2
(e.g., occupation)

A   C   B   A   C   B

Figure by MIT OCW.

Assumptions in
2ndG Navigation/Search Models III

- 1.   Individuals have ***links and identities***

- ***2.***   Individuals partition the world into a layered hierarchy and ***distance,*** x_{ij} is assumed to be the height of the lowest common parent. The branching ratio, b, and levels, *l* define this abstraction

- 3.   Group membership signifies not only identity but also is a primary basis for determining social interaction: $p_x = c \exp[-\alpha x]$

- 4.   Individuals hierarchically partition the world in more than one way and the model first assumes these distinctions are independent

- 5.   Individuals construct a measure of "social distance" which is the minimum over all dimensions between the nodes (violates the triangle inequality)
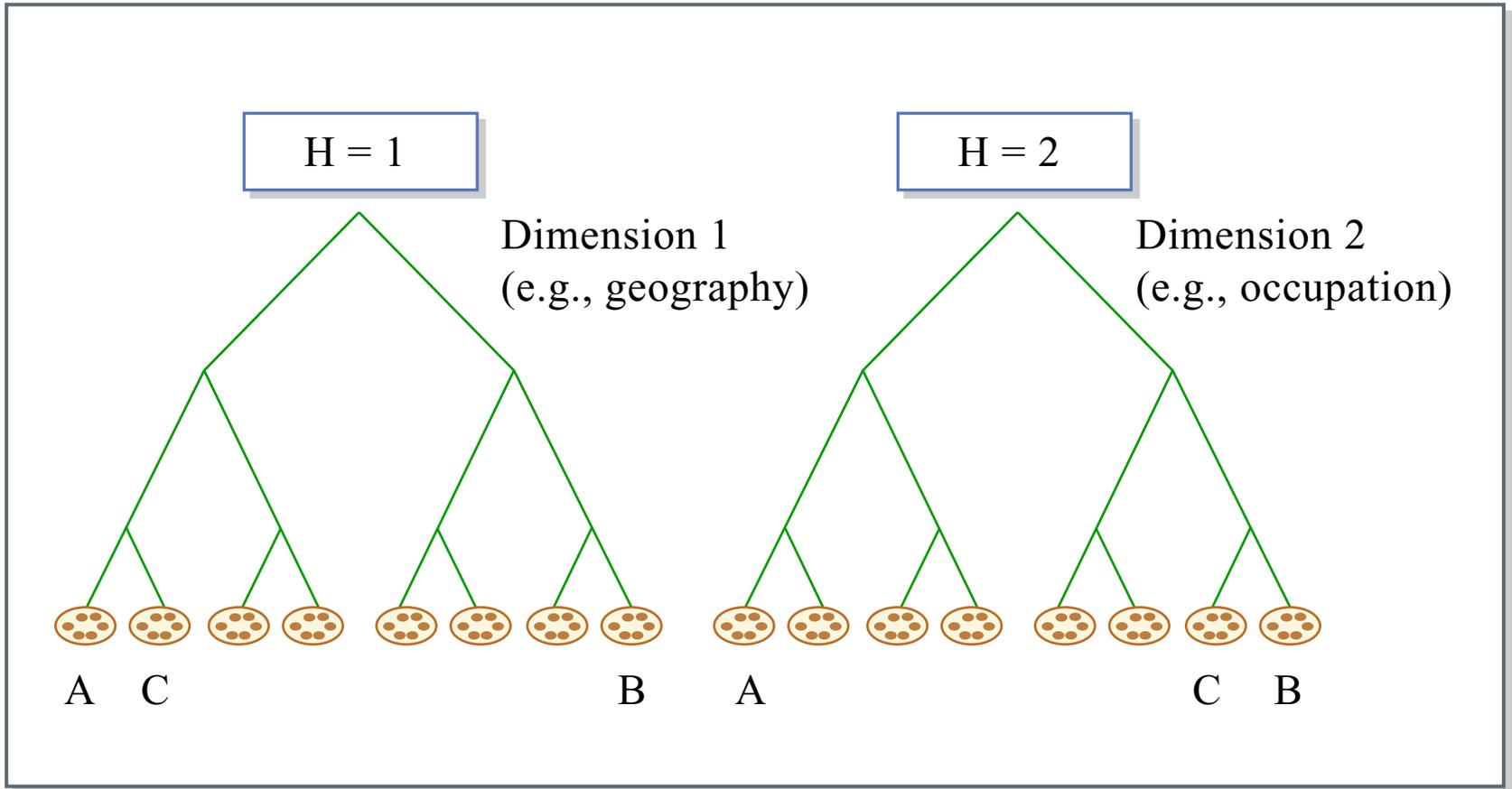
- 6.   Individuals forward messages based only on knowledge of their nearest neighbors and their identities. Forward the message to someone closer to the target is the "greedy" or "gossip" algorithm used

Results I

- Successful search assumes a decent probability (.05) of finishing the chain even though the probability of terminating the search *at each step* is fairly high (0.25 or higher)

- Key result is that searchable networks occupy a broad range of parameter space (α, H) with almost all searchable networks having α >0 and H >1

Massachusetts Institute of Technology

Search success for different size networks with alpha and H

Figure by MIT OCW.

Massachusetts Institute of Technology
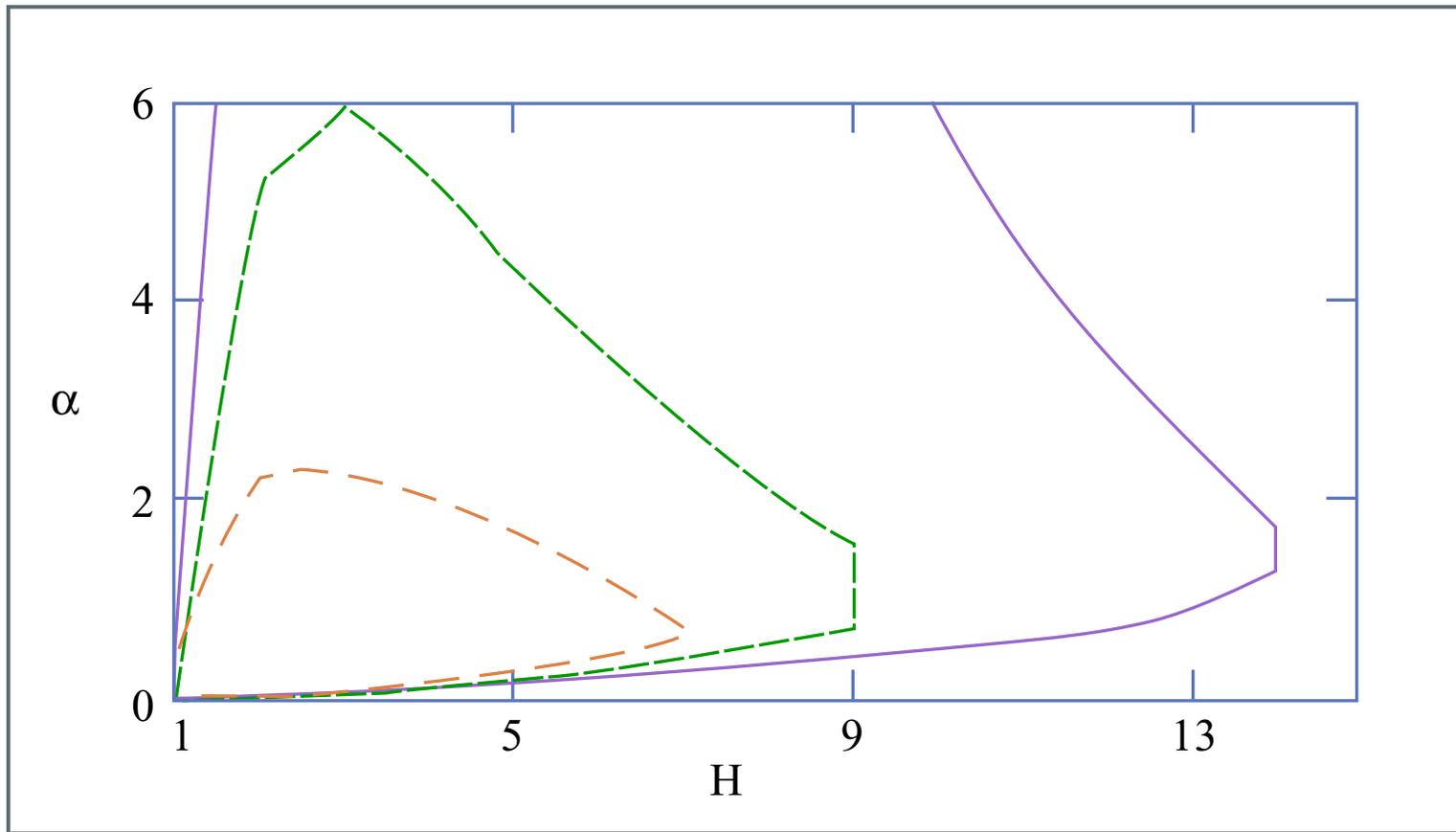
Results II

- Successful search assumes a decent probability (.05) of finishing the chain even though the probability of terminating the search *at each step* is fairly high (0.25 or higher)

- Key result is that searchable networks occupy a broad range of parameter space (α , H) with almost all searchable networks having α >0 and H >1

- Increasing group dimension beyond H = 1 yields a dramatic increase in search success (= reduction in delivery time) but "the improvement is lost as H increases further"
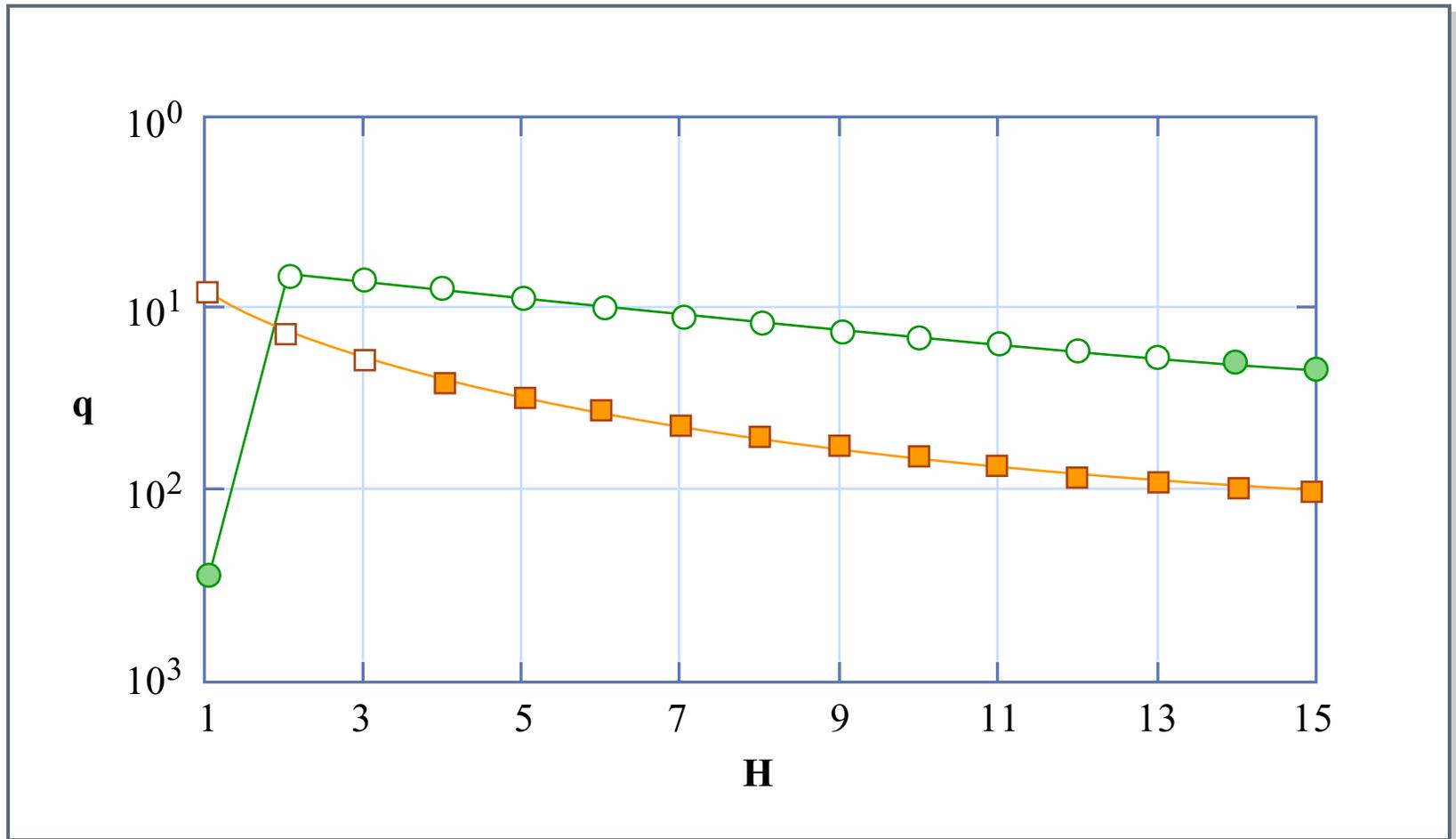
Probability of successful search

Figure by MIT OCW.

Massachusetts Institute of Technology

Results III

- Successful search assumes a decent probability (.05) of finishing the chain even though the probability of terminating the search *at each step* is fairly high (0.25 or higher)

- Key result is that searchable networks occupy a broad range of parameter space (α, H) with almost all searchable networks having α >0 and H >1

- Increasing group dimension beyond H = 1 yields a dramatic increase in search success (= reduction in delivery time) but the improvement is lost as H increases further

- For plausible values of all parameters, agreement with Milgram results are found
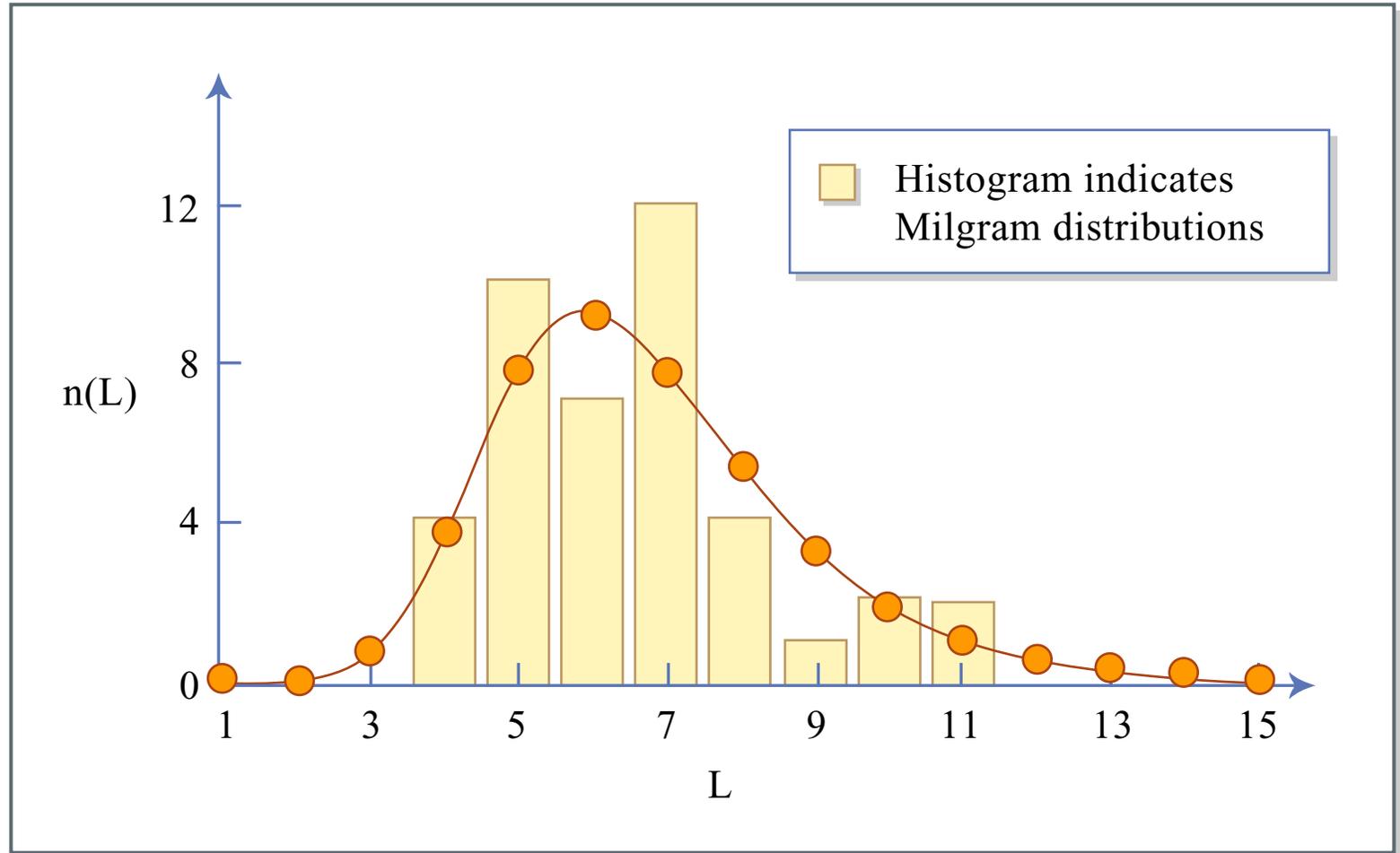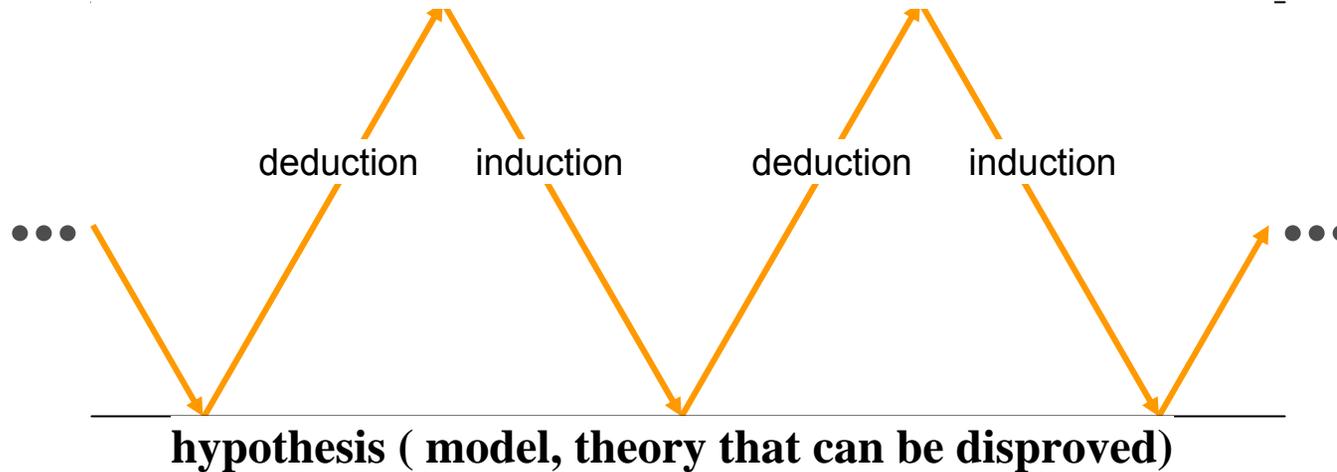
Distribution predicted vs. Milgram distributions

Histogram indicates Milgram distributions

Figure by MIT OCW.

Massachusetts Institute of Technology

The Iterative Learning Process

Objectively obtained quantitative data (facts, phenomena)

deduction   induction   deduction   induction

hypothesis (model, theory that can be disproved)

Models are "hardened" only by intensive simultaneous observational studies of relevant **reality**.

●●● What social distance (communication) exists in real social networks?
Random network models indicate relatively short paths might exist.
Milgram does an experiment and short paths (small worlds) exist.
Random networks do not describe clustering and short paths
Small world model is consistent and ubiquitous- Milgram experiment is revisited
Kleinberg points out navigation issue and introduces a model which treats it
but does not agree with Milgram. A 2nd generation navigation model
introduces structure into the social network and agrees with Milgram result.

Possible Future Research - Sociological Network Models

- Refinement of sociological network models
  - Clear measurement of identity hierarchies
  - Add strength of ties by methodology Newman has developed for weighted networks
- Collaborative Problem Solving in Large Organizations
  - **Community** by knowledge area vs. collaboration by problem content vs. collaboration by previous success
- Collaboration by Internet (WWW)
  - Social Identity Hierarchy vs. non-internet
  - Interest Groups vs. age and economics
- Social Networks within organizational hierarchies
  - Identification of important characteristics that determine such networks (age, hiring group, educational institution, neighborhood, functional specialty, co-workers, etc.) and possible role/utility in organizational architecture and effectiveness
  - Influence on conflict and cooperation in organizations

Possible Future Research and Applications of Sociological Network Models b.

- Marketing Research
    - How congruent are groupings that are made in marketing research with the social network communities?
    - Can one use known communication and search results to design more effective marketing/advertising strategies?
- Stakeholder Analysis
    - Should we think of stakeholders as part of a larger sociological network?
    - What relationships exist among different stakeholders

- Study of sociological networks over time
    - Permanence of identity
    - Influence of communication technology on identity

Modeling Overview

- Model Hardening (evolution with fitness checks)
- Model types in framework
    - System Formation (constraints-structure) Models
    - Structure-Property (System behavior) Models
    - System Observation Models
- Model Limitations and Benefits- general discussion
- System types of application interest
- System Architecting Purposes

Modeling (limitations and benefits)

- We are spending a good deal of time (and will spend more) on quantitative (mathematical) models for complex systems based on various "network approximations"
- As an ***introduction*** to the subject of how good the "network approximations" should be "before we believe" them, let's discuss the "goodness" of mathematical models generally:
  - Limitations and concerns
  - Value and possible benefits

How can models (and attempts to model) be harmful?

- If they are used thoughtlessly
  - A convenient metric becomes confused with "goodness"
  - Other trusted indicators are ignored because the model indicates otherwise

- If the attempt to be quantitative makes one deliberately eliminate important aspects of the problem.
  - I (perhaps unfairly) call this the "OR failure" but many avoid social complexity because of distaste with modeling humans or groups of humans. Many (all?) of *our important problems* involve human and social complexity.

Modeling (limitations and benefits) II

- We are spending a good deal of time (and will spend more) on quantitative (mathematical) models for complex systems based on various "network approximations"
- As an ***introduction*** to the subject of how good the "network approximations" should be "before we believe" them, let's discuss the "goodness" of mathematical models generally:
  - Limitations and concerns
  - Value and possible benefits

- "**All models are *wrong* but some are more *useful* than others**" is from George Box
  - J. D. C. Little wrote an earlier well-known article about this topic (without using the quote).

Massachusetts Institute of Technology

Why are all Models Wrong? Some possibilities

- Right model → Inaccurate answer
  - Rounding error
  - Truncation error
  - **I**ll conditioning
- Right model → Misleading answer
  - Chaotic systems
- Right model → No answer whatsoever
  - **Failure to converge**
  - **Algorithmic complexity**
- Not-so right model → Inaccurate answer
  - **Unmodeled effects**
  - Bugs in coding the model

**The Fundamental
Issues-*especially
for complex systems***

Massachusetts Institute of Technology

NP-Complete

- The class <u>NP-Complete</u> is a set of problems that, we believe, has no polynomial time algorithms.
- Therefore, they are hard problems.
- If a problem is NP-complete, there is no hope to solve the problem efficiently when n becomes large.

Computational Complexity and Moore's Law

- Consider a problem that requires 3^n flops
- World's fastest computer ~ 36 Teraflops/sec
- In a week, you can solve a problem where
$$n=\log(60*60*24*7*36*10^{12})/\log(3)=40$$
- If Moore's Law continues for 10 more years
$$n=\log(2^{10/1.5}*60*60*24*7*36*10^{12})/\log(3)=44$$

Our problems-if described in full detail- can have $n > 10^8$

In what senses are models useful?

Why Models can be Useful

- Engineering/managing
    - Estimating design parameters
    - Sensitivity of properties (performance attributes, cost, constraints, ilities, etc) to design parameters
    - Quantifying tradeoffs among attributes
    - Exploring alternate architectures
    - Knowing what to test and how to test it
    - *Wide applicability and testability when "first principles" are understood*

- Science
    - *Testability* and predictive power
    - To suggest new experiments
    - Powerful *condensation of knowledge*
    - Transferable to apparently different domains
    - The "essence" of a phenomenon can be clearly understood
    - New qualitative insights
    - *Stable "Principles" can emerge* (and re-emerge thus watch "first principles" meaning)

My Bottom Line is that use of models in Engineering and Science are very closely related

Massachusetts Institute of Technology

The fundamental modeling tradeoff

- The second important modeling quote
  - "A model (or theory) should be *as simple as possible* and not *any* simpler"

Massachusetts Institute of Technology

Modeling Overview II

- Model Hardening (evolution with fitness checks)
- Model types in framework
    - System Formation (constraints-structure) Models
    - Structure-Property (System behavior) Models
- Model Limitations and Benefits- general discussion
- System types of application interest. What complex systems do we want to model?
    - Sociological
        - Informal or self-organized
        - Formal or organized by design
    - Technological (and biological)

- System Architecting Purposes

Massachusetts Institute of Technology

Modeling Overview III

- Model Hardening (evolution with fitness checks)
- Model types in framework
    - System Formation (constraints-structure) Models
    - Structure-Property (System behavior) Models
- Model Limitations and Benefits- general discussion
- System types of application interest
- System Architecting Purposes
    - Organizational "goodness" (designed sociological network-to *design more effective organizations*)
    - Robustness, Flexibility (and other properties?) in "Engineering Systems" or systems that simultaneously possess *high levels of social and technical complexity*
    - *To understand how to choose (and cause to occur) better architectures for such systems*
- *Discussion of 2nd generation Navigation Model*
    - Moderately hardened,
    - Simplicity, practical use, other benefits…

Massachusetts Institute of Technology

References for Lecture 14

- Chapter 5 in *Six Degrees*
- J. Kleinberg, "Navigation in a Small World", *Nature*, (2000)
- J. Kleinberg, "Small-World Phenomena and the Dynamics of Information", *NIPS,* (2001)

- Watts,D. J., Dodds, P. S. and M. E. J. Newman, "Identity and Search in Social Networks" *Science,* 296 (May 2002).

- Little, John D. C (April 1970). "Models and Managers: The Concept of A Decision Calculus." *Management Science* 16(8):B466-B485.