Lecture 28: Course summary

INFO 2950:
Mathematical Methods for
Information Science

One last application: finding ‘bursts’ in email

Idea: one way to organize email is to divide it into periods (by date) in which some term occurs frequently and when it occurs infrequently.

E.g. ‘prelim’

How could we automatically detect such ‘bursts’?

Idea #1

Suppose we divide email into days, and check how many emails contain the term in a day.

ave. λ = # of emails containing the term in a day

What might be a reasonable probability distribution for the number of emails containing the term in a day?

Poisson distribution $Pr[k \text{ things}] = \frac{\lambda^k e^{-\lambda}}{k!}$

We will need two distributions, one for when the term is getting mentioned a lot (the burst) and one when it is not mentioned so much.

$\lambda_1 << \lambda_2$

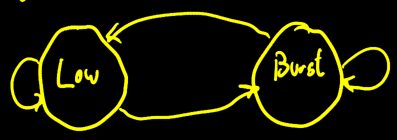
What should be true of the two distributions?

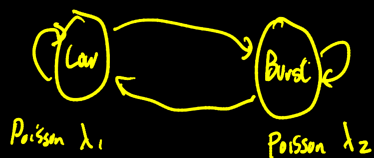
"Low" period $\frac{\lambda_1^k e^{-\lambda_1}}{k!}$ "Burst" $\frac{\lambda_2^k e^{-\lambda_2}}{k!}$

Idea #2

How can we model the transition(s) between when we have a burst and when we don't?

Markov Chain





Now given a bunch of email, how can we figure out when a burst was most likely occurring?

Use Hidden Markov Model

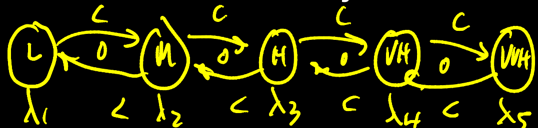
Viterbi's alg to decide which days 'Low'

days 'Burst'

This is (mostly) the idea of a paper of Kleinberg, "Bursts and Hierarchical Structure in Streams" (2002).

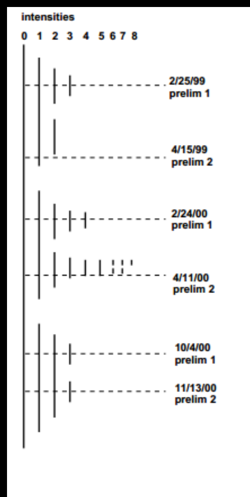
Two changes:

- Have lots of states, not just two

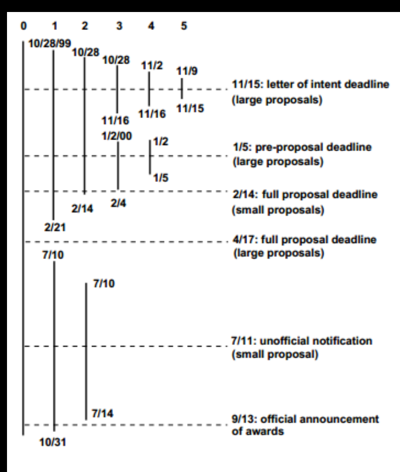


- Use Viterbi to minimize cost, not maximize probability

Results for ‘prelim’



Result for ‘ITR’



Applying to texts

Can apply to all the words in a body of texts and see when those words have bursts.

Example: Presidential State of the Union addresses.

Another example: titles from the database research community.

Where have we been?

What will we do?

Introduction, set theory

Probability and statistics (~8 lectures)

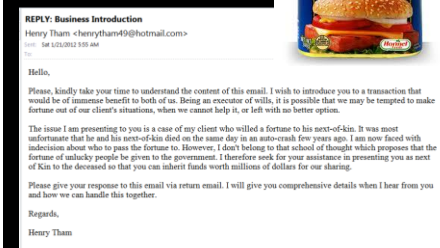
Graph theory and algorithms (~8 lectures)

Markov models and algorithms (~6 lectures)

Finite state automata (~4 lectures)

How to detect spam

How does a computer know when a message is spam?



REPLY: Business Introduction

Henry Tham <henrytham49@hotmail.com>

Sent: Sat 1/21/2012 5:55 AM

To:

Hello,

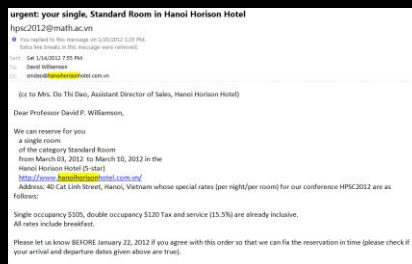
Please, kindly take your time to understand the content of this email. I wish to introduce you to a transaction that would be of immense benefit to both of us. Being an executor of wills, it is possible that we may be tempted to make fortune out of our client's situations, when we cannot help it, or left with no better option.

The issue I am presenting to you is a case of my client who willed a fortune to his next-of-kin. It was most unfortunate that he and his next-of-kin died on the same day in an auto-crash few years ago. I am now faced with indecision about who to pass the fortune to. However, I don't belong to that school of thought which proposes that the fortune of unlucky people be given to the government. I therefore seek for your assistance in presenting you as next of Kin to the deceased so that you can inherit funds worth millions of dollars for our sharing.

Please give your response to this email via return email. I will give you comprehensive details when I hear from you and how we can handle this together.

Regards,

Henry Tham



urgent: your single, Standard Room in Hanoi Horison Hotel

hpsc2012@math.ac.vn

You replied to this message on 1/20/2012 1:25 PM.
Extra line breaks in this message were removed.

Sent: Sat 1/14/2012 7:55 PM

To: David Williamson

Cc: smdao@hanoihorisonhotel.com.vn

(cc to Mrs. Do Thi Dao, Assistant Director of Sales, Hanoi Horison Hotel)

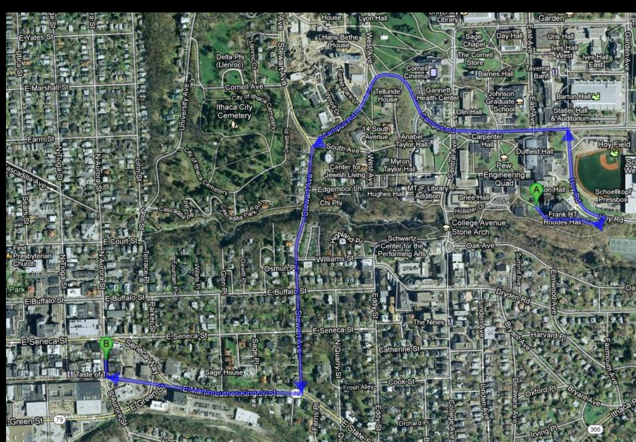
Dear Professor David P. Williamson,

We can reserve for you
a single room
of the category Standard Room
from March 03, 2012 to March 10, 2012 in the
Hanoi Horison Hotel (5-star)
http://www.hanoihorisonhotel.com.vn/
Address: 40 Cat Linh Street, Hanoi, Vietnam whose special rates (per night/per room) for our conference HPSC2012 are as follows:

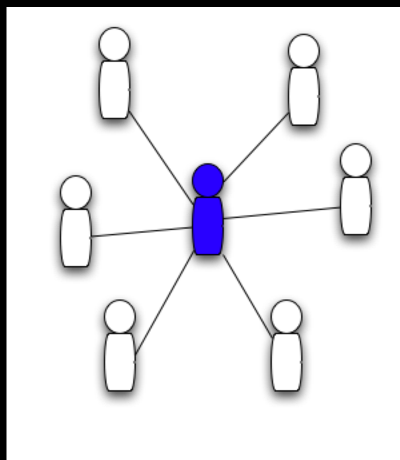
Single occupancy $105, double occupancy $120 Tax and service (15.5%) are already inclusive.
All rates include breakfast.

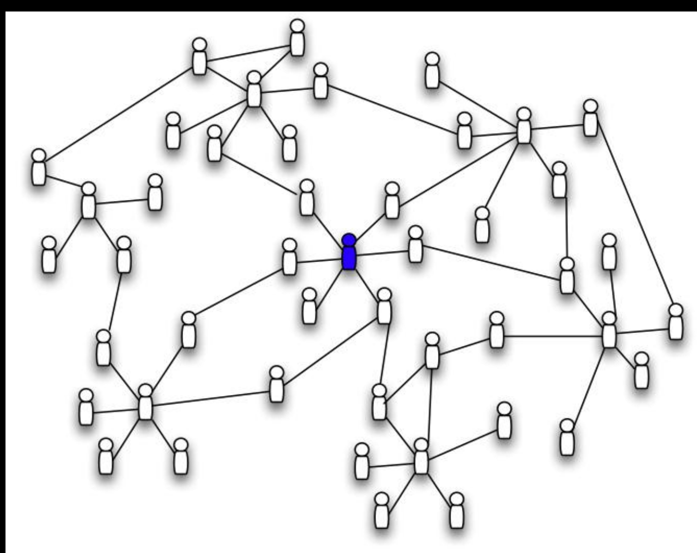
Please let us know BEFORE January 22, 2012 if you agree with this order so that we can fix the reservation in time (please check if your arrival and departure dates given above are true).

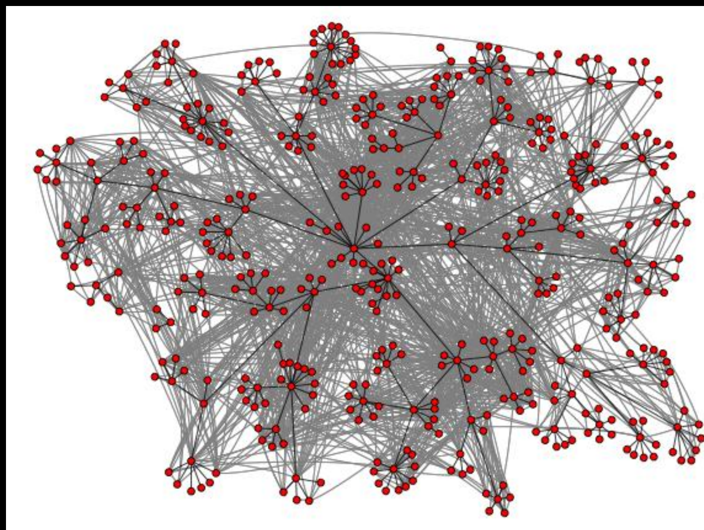
How does Google Maps work to find directions?



What can we do with social networks?

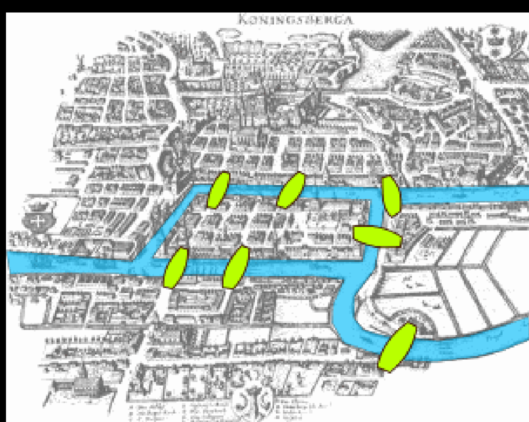




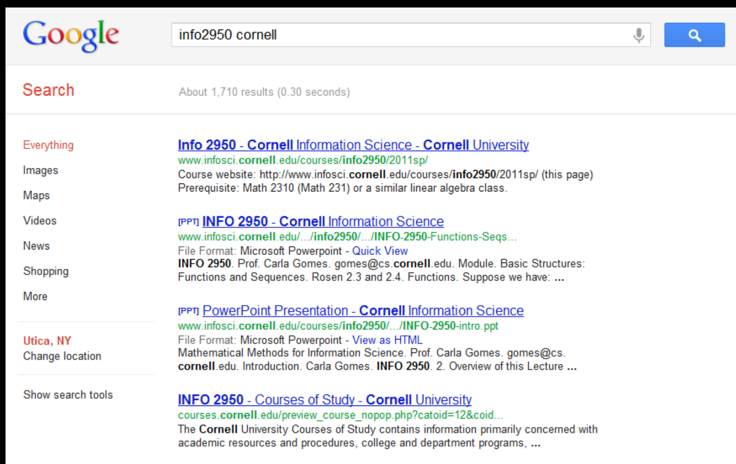


Adamic 05

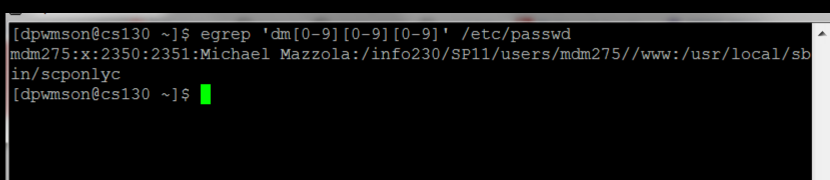
And what does it have to do with bridges in Königsberg?



How does Google find the page you want?



And how can you quickly find the information you want?



About the final

Wednesday, May 16, 2-4:30PM, Hollister 401

Open book, open notes: any notes you took yourself, any material from the course website, and Rosen

Comprehensive

Coverage

- Set theory
  - Basic terminology
  - Set operations (incl. Cartesian product, power set)
- Probability
  - Finite probability spaces, events
  - Counting and ordering; binomial coefficients and factorials
  - Uniform probability distributions
  - Tricks for computing probability
  - Joint probability, conditional probability
  - Bayes' theorem
    - Bayesian spam filtering
    - Naïve Bayes assumptions
  - Random variables
    - Expected values, variance, standard deviation
  - Bernoulli trials/binomial distribution
  - Central limit theorem
  - Rare events and the Poisson distribution

- Exponentials and logarithms
  - Expressions for e
  - Manipulating exponentials and logarithms
- Graph theory
  - Basic terminology (incl. paths, cycles, trees)
  - Eulerian paths and Eulerian circuits
    - Conditions under which these exist
  - Hamiltonian circuits
    - Conditions under which these exist
    - NP-complete (rough definition)
  - Planar graphs
    - Euler's formula and consequences
    - Kuratowski's theorem
  - Graph coloring (incl. planar graphs)
  - Spanning trees
    - Finding a spanning tree (including depth-first and breadth-first search)
    - Minimum spanning trees (including Kruskal and Prim's algorithm)
  - Shortest paths and Dijkstra's algorithm
  - The traveling salesman problem
    - Why it is as hard as finding a Hamiltonian circuit
    - Finding a near-optimal tour
  - The small world phenomenon and random graphs
    - Milgram's experiment
    - Erdos-Renyi/Watts-Strogatz/Kleinberg random graphs

The web and PageRank
- A brief history of the web
- PageRank
- HITS
- Eigenvalues/eigenvectors

Markov chains
- Calculating probabilities and expectations
- Types of states and ergodic chains
- Stationary distributions
- Estimating transition probabilities
- Applications: Mark V. Shaney and speech recognition
- Hidden Markov models and the Viterbi algorithm

Finite automata
- Deterministic finite automata
- Nondeterministic finite automata
- Equivalence of nondeterministic and deterministic finite automata (the *subset* construction)
- Nonregular languages
- Regular expressions
- The equivalence of finite automata and regular expressions
- Turing machines, including the Church/Turing thesis and the halting problem

Practice final review

Monday May 14, 1-3, room TBA

Office hours

Tuesday May 8, 11-12

Wednesday May 9, 11-12

Friday May 11, 11-12

Tuesday May 15, 10-12

By appointment

Course evaluation

Fill out the College of Engineering course evaluation for INFO 2950, get 5 bonus points on the final.

If you want to know more…

Probability

ENGRD 2700
ORIE 3120 (databases + stats)
ORIE 3500 (requires ENGRD 2700)
MATH 4710 (requires calculus)
ORIE 4740 (data mining; requires 3500)

Graph theory and algorithms

CS 4820 Algorithms (needs CS 3110)

ORIE 4330 Discrete Models (needs ORIE 3300 and 2110)

Markov chains and applications

MATH 4740 (need 4710)

ORIE 3510 Stochastic Processes (need 3500)

CS 4780 Machine Learning (CS 2110)

Math models of computing

CS 4810 Theory of Computing

Best wishes for
finals and summer!