New Directions in Computer Science

John Hopcroft
Cornell University
Time of change

- The information age is a revolution that is changing all aspects of our lives.

- Those individuals, institutions, and nations who recognize this change and position themselves for the future will benefit enormously.
Computer Science is changing

Early years

- Programming languages
- Compilers
- Operating systems
- Algorithms
- Data bases

Emphasis on making computers useful
Computer Science is changing

The future years

- Tracking the flow of ideas in scientific literature
- Tracking evolution of communities in social networks
- Extracting information from unstructured data sources
- Processing massive data sets and streams
- Extracting signals from noise
- Dealing with high dimensional data and dimension reduction
- The field will become much more application oriented
Computer Science is changing

Drivers of change

- Merging of computing and communication
- The wealth of data available in digital form
- Networked devices and sensors
Implications for TCS

- Need to develop theory to support the new directions

- Update computer science education
This talk consists of three parts.

- A view of the future.
- The science base needed to support future activities.
- What does a science base look like?
Big data

- We generate 2.5 exabytes of data/day, $2.5 \times 10^{18}$.
- We broadcast 2 zetta bytes per day.
  approximately 174 newspapers per day for every person on the earth.
- Maybe 20 billion web pages
Facebook

845 million
monthly active users

2.7 billion
Likes & Comments per day

250 million
photos uploaded per day

100 billion
friendships
Higgs Boson

CERN's Large Hadron Collider generates hundreds of millions of particle collisions each second. Recording, storing and analyzing these vast amounts of collisions presents a massive data challenge because the collider produces roughly 20 million gigabytes of data each year.

1,000,000,000,000,000: The number of proton-proton collisions, a thousand trillion, analyzed by ATLAS and CMS experiments.

100,000: The number of CDs it would take to record all the data from the ATLAS detector per second, or a stack reaching 450 feet (137 meters) high every second; at this rate, the CD stack could reach the moon and back twice each year, according to CERN.

27: The number of CDs per minute it would take to hold the amount of data ATLAS actually records, since it only records data that shows signs of something new.

"Without the worldwide grid of computing this result would not have happened," said Rolf-Dieter Heuer, director general at CERN during a press conference. The computing power and the network that CERN uses is a very important part of the research, he added.
Current database tools are insufficient to capture, analyze, search, and visualize the size of data encountered today.
Theory to support new directions

- Large graphs
- Spectral analysis
- High dimensions and dimension reduction
- Clustering
- Collaborative filtering
- Extracting signal from noise
- Sparse vectors
Sparse vectors

There are a number of situations where sparse vectors are important.

- Tracking the flow of ideas in scientific literature
- Biological applications
- Signal processing

Purdue University
Sparse vectors in biology

plants

Genotype
Internal code

= =

Phenotype
Observables
Outward manifestation

Genotype
Internal code
Digitization of medical records

- Doctor – needs my entire medical record
- Insurance company – needs my last doctor visit, not my entire medical record
- Researcher – needs statistical information but no identifiable individual information

Relevant research – zero knowledge proofs, differential privacy
A zero knowledge proof of a statement is a proof that the statement is true without providing you any other information.
Zero knowledge proof for Sudoku
Zero knowledge proof

- Graph 3-colorability

- Problem is NP-hard - No polynomial time algorithm unless P=NP
I send the sealed envelopes.
You select an edge and open the two envelopes corresponding to the end points.
Then we destroy all envelopes and start over, but I permute the colors and then resend the envelopes.
Digitization of medical records is not the only system

- Car and road – gps – privacy
- Supply chains
- Transportation systems
In the past, sociologists could study groups of a few thousand individuals.

Today, with social networks, we can study interaction among millions of individuals.

One important activity is how communities form and evolve.
Early work

- Min cut – two equal sized communities
- Conductance – minimizes cross edges

Future work

- Consider communities with more external edges than internal edges
- Find small communities
- Track communities over time
- Develop appropriate definitions for communities
- Understand the structure of different types of social networks
Our view of a community

Classmates

Family and friends

Me

Colleagues at Cornell

TCS

More connections outside than inside
Ongoing research on finding communities
Spectral clustering with K-means.
Spectral clustering with K-means.

Purdue University
Spectral clustering with K-means
What if communities overlap?

\[
\begin{array}{cc}
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 1 \\
1 & 1 \\
0 & 1 \\
0 & 1 \\
0 & 1 \\
\end{array}
\]
Instead of two overlapping clusters, we find three clusters.
Instead of clustering the rows of the singular vectors, find the minimum 0-norm vector in the space spanned by the singular vectors.

The minimum 0-norm vector is, of course, the all zero vector, so we will require one component to be 1.
Finding the minimum 0-norm vector is NP-hard.

Use the minimum 1-norm vector as a proxy. This is a linear programming problem.
What we have described is how to find global structure.

We would like to apply these ideas to find local structure.
We want to find community of size 50 in a network of size $10^9$. 
Instead of finding singular vectors, take a small number of steps in a random walk.

Look at early convergence of the random walk.

Find the minimum one norm vector in $A^5[x, Ax, A^2x]$.
Minimum 1-norm vector is not an indicator vector.

By thresh-holding the components, convert it to an indicator vector for the community.
Actually allow vector to be close to subspace.

\[ \min (|y|_1 + \tau \cos \theta) \]

↑
angle with subspace
Random walk

How long?

What dimension?
Structure of communities

- How many communities is a person in?
  Small, medium, large?
- How many seed points are needed to uniquely specify a community a person is in?
- Which seeds are good seeds?
- Etc.
What types of communities are there?

How do communities evolve over time?
Theory of Large Graphs

- Large graphs with billions of vertices
- Exact edges present not critical
- Invariant to small changes in definition
- Must be able to prove basic theorems
Erdős-Renyi

- $n$ vertices
- each of $n^2$ potential edges is present with independent probability

$$\binom{N}{n} p^n (1-p)^{N-n}$$

number of vertices

vertex degree

binomial degree distribution
Generative models for graphs

- Vertices and edges added at each unit of time
- Rule to determine where to place edges
  - Uniform probability
  - Preferential attachment - gives rise to power law degree distributions
Preferential attachment gives rise to the power law degree distribution common in many graphs.
Protein interactions

2730 proteins in data base
3602 interactions between proteins

<table>
<thead>
<tr>
<th>SIZE OF COMPONENT</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>…</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMBER OF COMPONENTS</td>
<td>48</td>
<td>179</td>
<td>50</td>
<td>25</td>
<td>14</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Only 899 proteins in components. Where are the 1851 missing proteins?

Science 1999 July 30; 285:751-753
Protein interactions

2730 proteins in data base
3602 interactions between proteins

<table>
<thead>
<tr>
<th>SIZE OF COMPONENT</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>...</th>
<th>1851</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMBER OF COMPONENTS</td>
<td>48</td>
<td>179</td>
<td>50</td>
<td>25</td>
<td>14</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Science 1999 July 30; 285:751-753

Purdue University
Science Base

What do we mean by science base?

- Example: High dimensions
High dimension is fundamentally different from 2 or 3 dimensional space
High dimensional data is inherently unstable.

Given $n$ random points in $d$-dimensional space, essentially all $n^2$ distances are equal.

\[ |x - y|^2 = \sum_{i=1}^{d} (x_i - y_i)^2 \]
High Dimensions

Intuition from two and three dimensions is not valid for high dimensions.

Volume of cube is one in all dimensions. Volume of sphere goes to zero.
Gaussian distribution

Probability mass concentrated between dotted lines
Gaussian in high dimensions

\[ \sqrt{d} \]

3
Two Gaussians

\[ \sqrt{d} \]

Purdue University
2 Gaussians with 1000 points each: mu=1.000, sigma=2.000, dim=500
2 Gaussians with 1000 points each: \( \mu=1.000, \sigma=2.000, \text{dim}=500 \)
Distance between two random points from same Gaussian

- Points on thin annulus of radius $\sqrt{d}$
- Approximate by a sphere of radius $\sqrt{d}$
- Average distance between two points is $\sqrt{2d}$

(Place one point at N. Pole, the other point at random. Almost surely, the second point will be near the equator.)
Expected distance between points from two Gaussians separated by $\delta$

$$\sqrt{\delta^2 + 2d}$$
Can separate points from two Gaussians if

\[ \sqrt{\delta^2 + 2d} > \sqrt{2d} + \gamma \]

\[ \sqrt{2d} \left(1 + \frac{1}{2} \frac{\delta^2}{2d} + \cdots \right) > \sqrt{2d} + \gamma \]

\[ \frac{1}{2} \frac{\delta^2}{\sqrt{2d}} > \gamma \]

\[ \delta > \sqrt{2\gamma} \left(2d\right)^{\frac{1}{4}} \]
Dimension reduction

- Project points onto subspace containing centers of Gaussians.

- Reduce dimension from $d$ to $k$, the number of Gaussians
Centers retain separation

Average distance between points reduced

by

\[ \sqrt{\frac{d}{k}} \]

\[
\begin{align*}
(x_1, x_2, \ldots, x_d) & \rightarrow (x_1, x_2, \ldots, x_k, 0, \ldots, 0) \\
d \bar{x}_i & \rightarrow k \bar{x}_i
\end{align*}
\]
Can separate Gaussians provided

$$\sqrt{\delta^2 + 2k} > \sqrt{2k} + \gamma$$

$$\delta > \text{some constant involving k and } \gamma$$

independent of the dimension
We have just seen what a science base for high dimensional data might look like.

For what other areas do we need a science base?
Ranking is important
- Restaurants, movies, books, web pages
- Multi-billion dollar industry

Collaborative filtering
- When a customer buys a product, what else is he or she likely to buy?

Dimension reduction

Extracting information from large data sources

Social networks
This is an exciting time for computer science.

There is a wealth of data in digital format, information from sensors, and social networks to explore.

It is important to develop the science base to support these activities.
Remember that institutions, nations, and individuals who position themselves for the future will benefit immensely.

Thank You!