NANO: Network Access Neutrality Observatory

Mukarram Bin Tariq, Murtaza Motiwala, Nick Feamster, Mostafa Ammar

Georgia Tech
CEO Edward Whitacre talks about the AT&T Wireless acquisition and how he's moving to keep abreast of cable competitors
CEO Edward Whitacre talks about the AT&T Wireless acquisition and how he’s moving to keep abreast of cable competitors

How concerned are you about Internet upstarts like Google (GOOG), MSN, Vonage, and others?
How do you think they're going to get to customers? Through a broadband pipe. Cable companies have them. We have them. Now what they would like to do is use my pipes free, but I ain't going to let them do that because we have spent this capital and we have to have a return on it. So there's going to have to be some mechanism for these people who use these pipes to pay for the portion they're using. Why should they be allowed to use my pipes?

The Internet can't be free in that sense, because we and the cable companies have made an investment and for a Google or Yahoo! (YHOO) or Vonage or anybody to expect to use these pipes [for] free is nuts!
Net Neutrality

CEO Edward Whitacre talks about the AT&T Wireless acquisition and how he’s moving to keep abreast of cable competitors

How concerned are you about Internet upstarts like Google (GOOG), MSN, Vonage, and others?

How do you think they’re going to get to customers? Through a broadband pipe. Cable companies have them. We have them. Now what they would like to do is use my pipes free, but I ain't going to let them do that because we have spent this capital and we have to have a return on it. So there's going to have to be some mechanism for these people who use these pipes to pay for the portion they’re using. Why should they be allowed to use my pipes?

The Internet can't be free in that sense, because we and the cable companies have made an investment and for a Google or Yahoo! (YHOO) or Vonage or anybody to expect to use these pipes [for] free is nuts!
Example: BitTorrent Blocking

It's Comcastic: Is Comcast Blocking Users From Seeding Torrents?

By Scott Gilbertson  August 20, 2007 | 12:48:27 PM  Categories: P2P

Bandwidth throttling and traffic shaping are nothing new, but rumors are making the rounds that Comcast is taking it to the next level when it comes to bittorrent traffic.

TorrentFreak claims that Comcast users are finding their torrent uploads throttled whenever they connect to non-Comcast users, which means you can’t seed torrents
Example: BitTorrent Blocking

http://broadband mpi-sws mpg.de/transparencce
Many Forms of Discrimination

Throttling and prioritizing based on destination or service

Target domains, applications, or content
Many Forms of Discrimination

Throttling and prioritizing based on destination or service

Target domains, applications, or content

Discriminatory peering

Resist peering with certain content providers

...
Problem Statement

Identify whether a degradation in a service performance is caused by discrimination by an ISP

Quantify the causal effect
Problem Statement

Identify whether a degradation in a service performance is caused by discrimination by an ISP

Quantify the causal effect

Existing techniques detect specific ISP methods

TCP RST (Glasnost)

ToS-bit based de-prioritization (NVLens)
Problem Statement

Identify whether a degradation in a service performance is caused by discrimination by an ISP

Quantify the causal effect

Existing techniques detect specific ISP methods

TCP RST (Glasnost)

ToS-bit based de-prioritization (NVLens)

**Goal:** Establish a *causal relationship* in the general case, without assuming anything about the ISP’s methods
Causality: An Analogy from Health

• Epidemiology: study causal relationships between risk factors and health outcome

• NANO: infer causal relationship between ISP and service performance
Does Aspirin Make You Healthy?
Does Aspirin Make You Healthy?

Sample of patients

Positive correlation in health and treatment

<table>
<thead>
<tr>
<th></th>
<th>Aspirin</th>
<th>No Aspirin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>40%</td>
<td>15%</td>
</tr>
<tr>
<td>Not Healthy</td>
<td>10%</td>
<td>35%</td>
</tr>
</tbody>
</table>
Does Aspirin Make You Healthy?

Sample of patients
Positive correlation in health and treatment
Can we say that Aspirin causes better health?

<table>
<thead>
<tr>
<th></th>
<th>Aspirin</th>
<th>No Aspirin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>40%</td>
<td>15%</td>
</tr>
<tr>
<td>Not Healthy</td>
<td>10%</td>
<td>35%</td>
</tr>
</tbody>
</table>
Does Aspirin Make You Healthy?

Sample of patients
Positive correlation in health and treatment
Can we say that Aspirin causes better health?

Confounding Variables: correlate with both cause and outcome variables and confuse the causal inference
Does an ISP Cause Service Degradation?
Does an ISP Cause Service Degradation?

Sample of client performances

Some correlation in ISP and service performance

<table>
<thead>
<tr>
<th>BitTorrent Download Time</th>
<th>Comcast</th>
<th>No Comcast</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5 sec</td>
<td>2 sec</td>
</tr>
</tbody>
</table>
Does an ISP Cause Service Degradation?

Sample of client performances

Some correlation in ISP and service performance

Can we say that Comcast is discriminating?

Many confounding variables can confuse the inference.

<table>
<thead>
<tr>
<th>BitTorrent Download Time</th>
<th>Comcast</th>
<th>No Comcast</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 sec</td>
<td>2 sec</td>
<td></td>
</tr>
</tbody>
</table>
Causation vs. Association (1)

Causal Effect = \( E(\text{Real Download time using Comcast}) - E(\text{Real Download time not using Comcast}) \)
Causation vs. Association (I)

Causal Effect = \[ E(\text{Real Download time using Comcast}) - E(\text{Real Download time not using Comcast}) \]

Performance with the ISP
Causation vs. Association (1)

Causal Effect = \[ \text{Performance with the ISP} - \text{Baseline Performance} \]

\[ E(\text{Real Download time using Comcast}) - E(\text{Real Download time not using Comcast}) \]
Causation vs. Association (I)

Causal Effect = $\theta = \mathbb{E}(G_1) - \mathbb{E}(G_0)$
$G_1, G_0$: Ground-truth values for performance
(aka. Counter-factual values)

Performance with the ISP
$\mathbb{E}(\text{Real Download time using Comcast}) - \mathbb{E}(\text{Real Download time not using Comcast})$
Baseline Performance
Causation vs. Association (I)

Causal Effect = \( \theta = \mathbb{E}(G_1) - \mathbb{E}(G_0) \)

\( G_1, G_0 \): Ground-truth values for performance
(aka. Counter-factual values)

Problem: Generally, we do not observe both ground truth values for the same clients.
Consequently, in situ data sets are not sufficient to directly estimate causal effect.
Causation vs. Association (2)

We can observe association in an \textit{in situ} data set.
Causation vs. Association (2)

We can observe association in an *in situ* data set.

Association = $E(\text{Download time using Comcast}) - E(\text{Download time not using Comcast})$
Causation vs. Association (2)

We can observe association in an *in situ* data set.

Association = \[ \text{Observed Performance with the ISP} \]
\[ E(\text{Download time using Comcast}) \]
\[ - E(\text{Download time not using Comcast}) \]
Causation vs. Association (2)

We can observe association in an *in situ* data set.

\[
\text{Association} = \frac{\text{Observed Performance with the ISP}}{\text{Observed Baseline Performance}}
\]

\[
\frac{\text{E(Download time using Comcast)}}{\text{E(Download time not using Comcast)}}
\]
Causation vs. Association (2)

We can observe association in an \textit{in situ} data set.

\[ \alpha = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0) \]

Association = \textit{Observed Performance with the ISP}

\begin{align*}
E(\text{Download time using Comcast}) - \\
E(\text{Download time not using Comcast})
\end{align*}

\textit{Observed Baseline Performance}
Causation vs. Association (2)

We can observe association in an *in situ* data set.

**Observed Performance with the ISP**

\[
\alpha = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0)
\]

In general, \( \alpha \neq \theta \).
Causation vs. Association (2)

We can observe association in an *in situ* data set.

\[
\text{Association} = E(\text{Download time using Comcast}) - E(\text{Download time not using Comcast})
\]

\[
\alpha = E(Y|X = 1) - E(Y|X = 0)
\]

*Observed Performance with the ISP*

In general, \( \alpha \neq \theta \).

How to estimate causal effect (\( \theta \))?
Estimating the Causal Effect

Two common approaches

a. Random Treatment

b. Adjusting for Confounding Variables
Random Treatment
Random Treatment

Given a population:
Random Treatment

Given a population:

1. Treat subjects with Aspirin randomly, irrespective of their health
Random Treatment

Given a population:

1. Treat subjects with Aspirin randomly, irrespective of their health
2. Observe new outcome and measure association

\[ \alpha = 0.8 - 0.25 = 0.55 \]
Random Treatment

Given a population:

1. Treat subjects with Aspirin randomly, irrespective of their health

2. Observe new outcome and measure association

3. For large samples, association converges to causal effect if confounding variables do not change

Diet, other drugs, etc. should not change

\[ \alpha = 0.8 - 0.25 = 0.55 \]
Random Treatment
(How to apply to the ISP Case?)
Random Treatment
(How to apply to the ISP Case?)

• Ask clients to change their ISP to an arbitrary one
Random Treatment
(How to apply to the ISP Case?)

• Ask clients to change their ISP to an arbitrary one

• Difficult to achieve on the Internet

  Changing ISP is cumbersome for the users

  Changing ISP may change other confounding variables, i.e., the ISP network
Adjusting for Confounding Variables

1. List confounders
   e.g., gender =\{○, □\}
Adjusting for Confounding Variables

1. List confounders
e.g., gender = { ○ , □ }

2. Collect a data set

An *in situ* data set

Treatment:
Baseline: Border(□, ○)
Treated: No border

Outcome:
Healthy (H), Not Healthy (!H)

Stratum: Type {Circle, Square}
Adjusting for Confounding Variables

1. List confounders
e.g., gender = {♀, ♂}

2. Collect a data set

3. Stratify along confounder variable values
Adjusting for Confounding Variables

1. List confounders
e.g., gender = {○, □}

2. Collect a data set

3. Stratify along confounder variable values
Adjusting for Confounding Variables

1. List confounders
   e.g., gender = { , }

2. Collect a data set

3. Stratify along confounder variable values

4. Measure Association

<table>
<thead>
<tr>
<th>Strata</th>
<th>Treated</th>
<th>Baseline</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.75</td>
<td>0.2</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>0.44</td>
<td>0.55</td>
<td>-0.11</td>
</tr>
</tbody>
</table>
Adjusting for Confounding Variables

1. List confounders e.g., gender = {Ω, □}
2. Collect a data set
3. Stratify along confounder variable values
4. Measure Association
5. If there still is association, then it must be causation

<table>
<thead>
<tr>
<th>Strata</th>
<th>Treated</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.75</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>0.44</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Effect 0.55  -0.11
Adjusting for Confounding
(How to Apply to the ISP Case?)

Challenges
Adjusting for Confounding
(How to Apply to the ISP Case?)

Challenges

What is the baseline?
Adjusting for Confounding
(How to Apply to the ISP Case?)

Challenges

What is the baseline?
What are the confounding variables?
Adjusting for Confounding
(How to Apply to the ISP Case?)

Challenges

What is the baseline?
What are the confounding variables?
Is the list of confounders sufficient?
Adjusting for Confounding
(How to Apply to the ISP Case?)

Challenges

What is the baseline?

What are the confounding variables?

Is the list of confounders sufficient?

How to collect the data?
Adjusting for Confounding
(How to Apply to the ISP Case?)

Challenges

What is the baseline?
What are the confounding variables?
Is the list of confounders sufficient?
How to collect the data?
Can we infer more than the effect? e.g., the discrimination criteria
Adjusting for Confounding
(How to Apply to the ISP Case?)

Challenges

What is the baseline?

What are the confounding variables?

Is the list of confounders sufficient?

How to collect the data?

Can we infer more than the effect? e.g., the discrimination criteria
What is the Baseline?
What is the Baseline?

Baseline: service performance when ISP is NOT used
What is the Baseline?

Baseline: service performance when ISP is NOT used

We need to use some ISP for comparison

What if the one we use is not neutral
What is the Baseline?

Baseline: service performance when ISP is NOT used

We need to use some ISP for comparison

What if the one we use is not neutral

Solutions:

a. Use average performance over all other ISPs
What is the Baseline?

Baseline: service performance when ISP is NOT used

We need to use some ISP for comparison

What if the one we use is not neutral

Solutions:

a. Use average performance over all other ISPs

b. Use a lab model
What is the Baseline?

Baseline: service performance when ISP is NOT used

We need to use some ISP for comparison

What if the one we use is not neutral

Solutions:

a. Use average performance over all other ISPs
b. Use a lab model
c. Use service providers’ model
Determine Confounding Variables

Using Domain Knowledge
Determine Confounding Variables

Using Domain Knowledge

Client Side

Client setup (Network Setup)

Application (Browser, BT Client, VoIP client)

Resources (Memory, CPU, Utilization)
Determine Confounding Variables

Using Domain Knowledge

Client Side

Client setup (Network Setup)
Application (Browser, BT Client, VoIP client)
Resources (Memory, CPU, Utilization)

ISP Related

Not all ISPs are equal; e.g., location.
Determine Confounding Variables

Using Domain Knowledge

Client Side

Client setup (Network Setup)

Application (Browser, BT Client, VoIP client)

Resources (Memory, CPU, Utilization)

ISP Related

Not all ISPs are equal; e.g., location.

Temporal

Diurnal cycles, transient failures
Inferring the Criteria
Inferring the Criteria

Label data in two classes:

discriminated (-),
non-discriminated (+)
Inferring the Criteria

Label data in two classes:
- discriminated (-),
- non-discriminated (+)

Train a decision tree for classification
Inferring the Criteria

Label data in two classes:

discriminated (-),
non-discriminated (+)

Train a decision tree for classification

rules provide hints about the criteria
A Simple Simulation
A Simple Simulation

• Clients use two applications: \( \text{App}_1 \) and \( \text{App}_2 \) to access services
• Service 1 is slower using \( \text{App}_2 \)
• App is confounding
• \( \text{ISP}_B \) throttles access to Service 1
A Simple Simulation

- Clients use two applications: App\textsubscript{1} and App\textsubscript{2} to access services.
- Service 1 is slower using App\textsubscript{2}.
- App is confounding.
- ISP\textsubscript{B} throttles access to Service 1.

<table>
<thead>
<tr>
<th>Association</th>
<th>Service 1</th>
<th>Service 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>7.68</td>
<td>2.67</td>
</tr>
<tr>
<td>ISP\textsubscript{B}</td>
<td>8.60</td>
<td>2.7</td>
</tr>
<tr>
<td>Association</td>
<td>0.92 (10%)</td>
<td>0.04 (1%)</td>
</tr>
</tbody>
</table>
A Simple Simulation

- Clients use two applications: App₁ and App₂ to access services
- Service 1 is slower using App₂
- App is confounding
- ISP_B throttles access to Service 1

<table>
<thead>
<tr>
<th>Association</th>
<th>Service 1</th>
<th>Service 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>7.68</td>
<td>2.67</td>
</tr>
<tr>
<td>ISP_B</td>
<td>8.60</td>
<td>2.7</td>
</tr>
<tr>
<td>Association</td>
<td><strong>0.92 (10%)</strong></td>
<td><strong>0.04 (1%)</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Causation</th>
<th>Service 1</th>
<th>Service 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>App₁: 9.90</td>
<td>App₂: 2.77</td>
</tr>
<tr>
<td>ISP_B</td>
<td>11.95</td>
<td>7.95</td>
</tr>
<tr>
<td>Causation</td>
<td><strong>2.05 (20%)</strong></td>
<td><strong>5.18 (187%)</strong></td>
</tr>
</tbody>
</table>
Conclusions
Conclusions

NANO: Black-box approach to infer and quantify discrimination; generically applicable.
Conclusions

NANO: Black-box approach to infer and quantify discrimination; generically applicable.

Ongoing work, Open Issues:

Privacy: Can we do local inference?
Conclusions

NANO: Black-box approach to infer and quantify discrimination; generically applicable.

Ongoing work, Open Issues:

Privacy: Can we do local inference?

Deployment: PlanetLab, CPR, Real Users
Conclusions

NANO: Black-box approach to infer and quantify discrimination; generically applicable.

Ongoing work, Open Issues:

Privacy: Can we do local inference?

Deployment: PlanetLab, CPR, Real Users

How much data?: Depends on variance
Conclusions

NANO: Black-box approach to infer and quantify discrimination; generically applicable.

Ongoing work, Open Issues:

Privacy: Can we do local inference?

Deployment: PlanetLab, CPR, Real Users

How much data?: Depends on variance

Contact: mmt@gatech.edu
Backup
Variables

• Causal Variables (X)
  The brand of an ISP
  IP address to ISP name mapping

• Outcome Variables (Y)
  Identify the Service
  The performance of a service
  Throughput, Delay, Jitter, Loss

• Confounding Variables (Z)
  Those that correlate with both
Adjusting for Confounding Variables

- Treatment (whether using ISP $i$): $X_i$
- Outcome (performance of service $j$): $Y_j$

Causal Effect:
Adjusting for Confounding Variables

• Treatment (whether using ISP \(i\)):\(X_i\)

Outcome (performance of service \(j\)):\(Y_j\)

Causal Effect: \(\theta_{ij}^{(z)} = \theta_{ij}(1; z) - \theta_{ij}(0; z)\)

\[
\theta_{ij} = \sum_z \theta_{ij}^{(z)}
\]

\[
\theta_{ij}(x; z) = \mathbb{E}(Y_j|X_i = x, Z = \mathbb{B}(z))
\]

\(\mathbb{B}(z)\) Boundaries of a stratum \(z\) with all-things equal
Sufficient Confounders?

If we have enough variables, then we should be able to predict the performance

Suppose: Confounding Var (Z), Treatment (X), Outcome (y)

Predict: \( \hat{y} = f(X; Z) \)

Test for prediction error: \( \frac{|y - \hat{y}|}{y} \)
Incentivizing

• Useful Diagnostics get a piggybacked ride

  NANO will identify transient failures in the elimination process

  Useful diagnostics dashboard for the users
P2P-ized Architecture

- Coordination to speed up inference and active learning