Today’s webinar topic is "Trustworthy Decision Making and Artificial Intelligence" with Arjan Durresi.
Our host is Jeannette Dopheide.

The meeting will begin shortly. Participants are muted. Click the chat button to ask a question.

This meeting will be recorded.

The Trusted CI Webinar Series is supported by National Science Foundation grant #1920430.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NSF.
Trustworthy Decision Making and Artificial Intelligence

Arjan Durresi
Department of Computer and Information Science
Indiana University - Purdue University Indianapolis, USA

www.cs.iupui.edu/~adurresi
adurresi@iupui.edu
Overview

- Why Trustworthy AI?
- Our Trust Management System
- Validation with Stock market data
- Trustability of Computing Systems
- Decision Making about Water Food Energy
- Detecting Fake users in Social Networks
- Huma-Machine in predicting Crimes
Computers in Decision Making

"All models are wrong, but some are useful."

George Box

“It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.”

Richard P. Feynman

“Understand the Model’s Errors before you understand the Model”

Nassim Taleb
Human-Machine Collaboration

- AI greater power is in complementing and *augmenting* human capabilities.
- Humans work with smart machines to exploit what each party does best.
- Humans, for example, are needed to develop, train, and manage various AI applications.
- In doing so, they are *enabling* those systems to function as true collaborative partners.
- For their part, machines are providing people with superhuman capabilities, such as the ability to process and analyze copious amounts of data from myriad sources in real time.
Many business, research, military, and other processes will take advantage of collaborative teams of humans working alongside machines.

As in all collaboration processes trust is a fundamental feature among the participants in the collaborative process as well as the trust on the final product of the process.

However, the collaboration system should recognize the specific capabilities of people and machines and enable the best combination of their capabilities.
Human - Machine Collaboration

- Training Machines
  - Select the right dataset for the given application

- Testing and explaining machines
  - AI, sophisticated algorithms can uncover numerous interesting correlations and lead to solutions that are not obvious.
  - Test algorithms and be responsible for holding any algorithm accountable for its results.
  - Selection of AI technology, such algorithms, and its features to be deployed for specific applications

- Controlling and overseeing machines

- Evaluation of the tradeoff between the risk and the trust level of an AI decision

- Guaranteeing the ethics compliance of machine actions
Trust

- Major social human feature
  - Refined by our evolution
  - Studied by many disciplines: psychology, sociology, neorophilosphy ...
    - Relate social functions to brain activities, especially in cortex
  - Essential in making decisions
Background and motivation

What is trust?

“Trust is the willingness of the trustor (evaluator) to take risk based on a subjective belief that a trustee (evaluatee) will exhibit reliable behavior to maximize the trustor's interest under uncertainty (e.g., ambiguity due to conflicting evidence and/or ignorance caused by complete lack of evidence) of a given situation based on the cognitive assessment of past experience with the trustee” [1]

Our Approach

- Treat trust similar to physical measurements
- Physical measurement accuracy can be improved by using more precise equipment, combining different measurement methods, or repeating the measurement
- Similarly, people develop an impression about another person based on interactions, experience.
Novelty and contributions

- Based on the similarity between trust assessment process and physical measurement, we propose a new measurement theory based trust management framework.

- Relate confidence with error in measurement theory, we are able to easily calculate trust propagation by using uncertainty/error propagation theory.

- Trust inference operators are ad-hoc in existing works, our framework can be adapted to various operators, i.e. multiplication, weighted mean, etc.

- We show the usage and validation of our trust management framework in two applications.
Trust Management System

Context

Decision Making

Trust Management

Trust Modelling
Trust metrics - trustworthiness

- Given a set of measurement results

\[ M = \{m_1, m_2, \ldots, m_k\} \text{ for each } m_i \in [0, 1] \]

- Trustworthiness or impression (m)
  - The trustor’s comprehensive summary of multiple “measurements” about the trustee’s trustworthiness, e.g. how good is Alice?
  - Similar to the mean of measurement results.

\[ m = \frac{\sum_{i=1}^{k} m_i}{k} \]
Trust metrics - confidence

- **Confidence (c)**
  - To what extent the trustor is certain/confident about the trust assessment.
  - Related with the standard error of the mean ($r$).

$$
r = \sqrt{\frac{\sum_{i=1}^{k} (m_i - m)^2}{k \times (k - 1)}}
$$

$$
c = \begin{cases} 
1 - 2 \cdot r, & \text{if } r \leq 0.5 \\
0, & \text{otherwise}
\end{cases}
$$

- Both trustworthiness ($m$) and confidence ($c$) of $T(m, c)$ are in the range of [0,1].
Confidence

- Given a fixed conflict ratio of evidence or measurements (i.e. positive vs. negative), confidence increases as the amount of evidence or measurements increases.
- Given a fixed amount of evidence or measurements, confidence increases when the conflict ratio decreases.
Uncertainty/error propagation theory

- Given two trust tuples T1(m1, c1) and T2(m2, c2), and a function of f(m1, m2). Propagated uncertainty or error can be calculated as following [2]:

\[ r_f^2 = \left( \frac{\partial f}{\partial m_1} \right)^2 r_1^2 + \left( \frac{\partial f}{\partial m_2} \right)^2 r_2^2 + 2 \frac{\partial f}{\partial m_1} \frac{\partial f}{\partial m_2} \text{cov}(m_1, m_2) \]

- By using the uncertainty propagation theory, our framework is flexible and can be adapted to various formulas.

Trust inference - transitivity

\[ A \rightarrow B \rightarrow Z \]

- A has a trust relationship with B (\( T^A_B \)), B has a trust relationship with Z (\( T^B_Z \)). To some extent, A also has a trust relationship with Z through B (\( T^{A:B}_Z \)).

\[ T^{A:B}_Z = T^A_B \otimes T^B_Z \]
Transitivity formulas

- **TP1 (Multiplication):**
  \[ m_B^A \otimes m_Z^B = m_B^A \times m_Z^B \]
  \[ r_B^A \otimes r_Z^B = \sqrt{(m_Z^B)^2 \times (r_B^A)^2 + (m_B^A)^2 \times (r_Z^B)^2} \]

- **TP2:**
  \[ m_B^A \otimes m_Z^B = m_B^A \times m_Z^B + (1 - m_B^A) \times (1 - m_Z^B) \]
  \[ r_B^A \otimes r_Z^B = \sqrt{(2 \times m_B^A - 1)^2 \times (r_Z^B)^2 + (2 \times m_Z^B - 1)^2 \times (r_B^A)^2} \]

- **TP3:**
  \[ m_B^A \otimes m_Z^B = m_{\text{min}} = \min(m_B^A, m_Z^B) \]
  \[ r_B^A \otimes r_Z^B = \min(r_i \text{ where } m_i = m_{\text{min}}) \]
Trust inference - aggregation

- A has trust relationships with B and C, B and C have trust relationship with Z. To some extent, A also has a trust relationship with Z through combining A and B's information.

\[ T_Z^{A:(B,C)} = T_Z^{A:B} \oplus T_Z^{A:C} \]
Aggregation formulas

- AP1 (averaging):
  \[ m_{Z:B}^A \oplus m_{Z:C}^A = \frac{m_{Z:B}^A + m_{Z:C}^A}{2} \]
  \[ r_{Z:B}^A \oplus r_{Z:C}^A = \sqrt{\frac{1}{2^2} \left( (r_{Z:B}^A)^2 + (r_{Z:C}^A)^2 \right)} \]

- AP2 (weighted mean):
  \[ m_{Z:B}^A \oplus m_{Z:C}^A = \frac{w_1 \cdot m_{Z:B}^A + w_2 \cdot m_{Z:C}^A}{\sum w_i} \]
  \[ r_{Z:B}^A \oplus r_{Z:C}^A = \sqrt{\frac{1}{(\sum w_i)^2} (w_1^2 \cdot (r_{Z:B}^A)^2 + w_2^2 \cdot (r_{Z:C}^A)^2)} \]

- AP3 (probability of the union of events):
  \[ m_{Z:B}^A \oplus m_{Z:C}^A = m_{Z:B}^A + m_{Z:C}^A - m_{Z:B}^A \cdot m_{Z:C}^A \]
  \[ r_{Z:B}^A \oplus r_{Z:C}^A = \sqrt{(1 - m_{Z:C}^A)^2 \cdot (r_{Z:B}^A)^2 + (1 - m_{Z:B}^A)^2 \cdot (r_{Z:C}^A)^2} \]
Aggregation formulas

- AP4 (strongest trustworthiness):

\[ m_{Z:B}^A \oplus m_{Z:C}^A = m_{max} = \max(m_{Z:B}^A, m_{Z:C}^A) \]

\[ r_{B}^{A:B} \oplus r_{Z}^{A:C} = \min(r_i \text{ where } m_i = m_{max}) \]

- AP5 (strongest confidence):

\[ r_{B}^{A:B} \oplus r_{Z}^{A:C} = r_{max}, \text{ where } c_{max} = \max(c_{Z:B}^A, c_{Z:C}^A) \]

\[ m_{Z}^{A:B} \oplus m_{Z}^{A:C} = \max(m_i \text{ where } c_i = c_{max}) \]
Validation experiments - datasets

- **Epinions.com dataset [3]:**
  
  Users are able to rate other users' reviews from un-useful (1 star) to very useful (5 stars).

- **Twitter dataset:**
  
  Collected a group of users who are followers of StockTwits in 2015.
  
  Collected all the users' tweets (up to 3,200 tweets for each user), only English tweets are included.

### Statistics of two datasets

<table>
<thead>
<tr>
<th></th>
<th>Epinions.com</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
<td>405,154</td>
<td>2,067,284</td>
</tr>
<tr>
<td>Number of ratings or Tweets</td>
<td>4,835,389</td>
<td>38,748,723</td>
</tr>
<tr>
<td>Number of links among users</td>
<td>1,530,103</td>
<td>6,829,998</td>
</tr>
</tbody>
</table>

Trust modeling for Epinions.com dataset

- Each rating is treated as a single measurement.

\[ m = \frac{\sum_{i=1}^{k} m_i}{k} \]

\[ r_r = \sqrt{\frac{\sum_{i=1}^{k} (m_i - m)^2}{k(k-1)}} \]

\[ r_s = \frac{\text{scale}}{2 \times \sqrt{3}} \]

\[ r = \sqrt{r_r^2 + r_s^2} \]

\[ c = \begin{cases} 1 - 2 \times r, & \text{if } r \leq 0.5 \\ 0, & \text{otherwise} \end{cases} \]
Trust modeling for Twitter dataset

- Interactive tweets: reply, mention, retweet, “@”.
- Sentiment analysis for interactive tweets (SentiStrength [4]) and normalization.
- Time windows: day, month, year. In each window (one month), the trustor has an impression on the trustee.

\[
m_{\text{month}} = \frac{\sum_{i=1}^{31} w_i m_i}{\sum_{i=1}^{31} w_i} \quad w_i = \frac{1}{r^{2.4}} \quad r_{\text{month}}^2 = \frac{1}{\sum_{i=1}^{31} w_i}
\]

- However, this impression becomes faded with time going on. So we introduce a forgetting factor, where is less than 1.

\[
c_i' = c_i \times \sigma^{12-i} \quad \sigma = 0.9
\]

- Combine all the windows' results using the weighted mean where weights are their confidence \(c_i'\).

---

Two hops results

MAE: mean absolute error

RMSE: root mean square error

\[ MAE(Man) = \frac{\sum_{i=1}^{n} |diff_{m_i}| + |diff_{c_i}|}{n} \]

Table II. Operators’ performances on the Epinions.com data set (two hops)

<table>
<thead>
<tr>
<th>Operators</th>
<th>MAE (diffm)</th>
<th>RMSE (diffm)</th>
<th>MAE (diffc)</th>
<th>RMSE (diffc)</th>
<th>MAE (Man)</th>
<th>RMSE (Man)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP1,AP1</td>
<td>0.0565</td>
<td>0.0827</td>
<td>0.1021</td>
<td>0.1170</td>
<td>0.1585</td>
<td>0.1433</td>
</tr>
<tr>
<td>TP1,AP2</td>
<td>0.0598</td>
<td>0.0924</td>
<td>0.0353</td>
<td>0.0630</td>
<td>0.0948</td>
<td>0.1118</td>
</tr>
<tr>
<td>TP1,AP3</td>
<td>0.0620</td>
<td>0.1250</td>
<td>0.1357</td>
<td>0.1476</td>
<td>0.1977</td>
<td>0.1934</td>
</tr>
<tr>
<td><strong>TP1,AP4</strong></td>
<td><strong>0.0511</strong></td>
<td><strong>0.1036</strong></td>
<td><strong>0.0514</strong></td>
<td><strong>0.0653</strong></td>
<td><strong>0.1026</strong></td>
<td><strong>0.1224</strong></td>
</tr>
<tr>
<td>TP1,AP5</td>
<td>0.1815</td>
<td>0.2497</td>
<td>0.0427</td>
<td>0.0614</td>
<td>0.2242</td>
<td>0.2572</td>
</tr>
<tr>
<td>TP2,AP1</td>
<td>0.0554</td>
<td>0.0808</td>
<td>0.1049</td>
<td>0.1197</td>
<td>0.1603</td>
<td>0.1444</td>
</tr>
<tr>
<td>TP2,AP2</td>
<td>0.0589</td>
<td>0.0906</td>
<td>0.0371</td>
<td>0.0659</td>
<td>0.0960</td>
<td>0.1121</td>
</tr>
<tr>
<td>TP2,AP3</td>
<td>0.0619</td>
<td>0.1248</td>
<td>0.1359</td>
<td>0.1476</td>
<td>0.1978</td>
<td>0.1933</td>
</tr>
<tr>
<td><strong>TP2,AP4</strong></td>
<td><strong>0.0510</strong></td>
<td><strong>0.1032</strong></td>
<td><strong>0.0501</strong></td>
<td><strong>0.0629</strong></td>
<td><strong>0.1012</strong></td>
<td><strong>0.1208</strong></td>
</tr>
<tr>
<td>TP2,AP5</td>
<td>0.2063</td>
<td>0.2656</td>
<td>0.0463</td>
<td>0.0705</td>
<td>0.2526</td>
<td>0.2748</td>
</tr>
<tr>
<td>TP3,AP1</td>
<td>0.0526</td>
<td>0.0783</td>
<td>0.1070</td>
<td>0.1215</td>
<td>0.1597</td>
<td>0.1446</td>
</tr>
<tr>
<td>TP3,AP2</td>
<td>0.0553</td>
<td>0.0870</td>
<td>0.0531</td>
<td>0.0750</td>
<td>0.1084</td>
<td>0.1149</td>
</tr>
<tr>
<td>TP3,AP3</td>
<td>0.0619</td>
<td>0.1249</td>
<td>0.1354</td>
<td>0.1478</td>
<td>0.1973</td>
<td>0.1935</td>
</tr>
<tr>
<td><strong>TP3,AP4</strong></td>
<td><strong>0.0513</strong></td>
<td><strong>0.1038</strong></td>
<td><strong>0.0526</strong></td>
<td><strong>0.0629</strong></td>
<td><strong>0.0778</strong></td>
<td><strong>0.1214</strong></td>
</tr>
<tr>
<td>TP3,AP5</td>
<td>0.0559</td>
<td>0.1106</td>
<td>0.0257</td>
<td>0.0621</td>
<td>0.0816</td>
<td>0.1268</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.4526</td>
<td>0.5346</td>
<td>0.3823</td>
<td>0.4643</td>
<td>0.8349</td>
<td>0.7081</td>
</tr>
</tbody>
</table>

Table III. Operators’ performances on the Twitter data set (two hops)

<table>
<thead>
<tr>
<th>Operators</th>
<th>MAE (diffm)</th>
<th>RMSE (diffm)</th>
<th>MAE (diffc)</th>
<th>RMSE (diffc)</th>
<th>MAE (Man)</th>
<th>RMSE (Man)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP1,AP1</td>
<td>0.2336</td>
<td>0.2527</td>
<td>0.1559</td>
<td>0.1946</td>
<td>0.3995</td>
<td>0.3189</td>
</tr>
<tr>
<td>TP1,AP2</td>
<td>0.2342</td>
<td>0.2533</td>
<td>0.1552</td>
<td>0.1937</td>
<td>0.3894</td>
<td>0.3189</td>
</tr>
<tr>
<td>TP1,AP3</td>
<td>0.2341</td>
<td>0.2575</td>
<td>0.1483</td>
<td>0.1893</td>
<td>0.3693</td>
<td>0.3083</td>
</tr>
<tr>
<td>TP1,AP4</td>
<td>0.2207</td>
<td>0.2409</td>
<td>0.1457</td>
<td>0.1800</td>
<td>0.3665</td>
<td>0.3008</td>
</tr>
<tr>
<td>TP1,AP5</td>
<td>0.2386</td>
<td>0.2590</td>
<td>0.1536</td>
<td>0.1923</td>
<td>0.3922</td>
<td>0.3226</td>
</tr>
<tr>
<td>TP2,AP1</td>
<td>0.1037</td>
<td>0.1456</td>
<td>0.2729</td>
<td>0.3250</td>
<td>0.3766</td>
<td>0.3562</td>
</tr>
<tr>
<td>TP2,AP2</td>
<td>0.1038</td>
<td>0.1457</td>
<td>0.2729</td>
<td>0.3252</td>
<td>0.3767</td>
<td>0.3563</td>
</tr>
<tr>
<td>TP2,AP3</td>
<td>0.1556</td>
<td>0.2198</td>
<td>0.2758</td>
<td>0.3270</td>
<td>0.4314</td>
<td>0.3948</td>
</tr>
<tr>
<td>TP2,AP4</td>
<td>0.1047</td>
<td>0.1468</td>
<td>0.2670</td>
<td>0.3192</td>
<td>0.3717</td>
<td>0.3514</td>
</tr>
<tr>
<td>TP2,AP5</td>
<td>0.1037</td>
<td>0.1460</td>
<td>0.2783</td>
<td>0.3305</td>
<td>0.3820</td>
<td>0.3613</td>
</tr>
<tr>
<td><strong>TP3,AP1</strong></td>
<td><strong>0.0692</strong></td>
<td><strong>0.1220</strong></td>
<td><strong>0.1224</strong></td>
<td><strong>0.2087</strong></td>
<td><strong>0.1916</strong></td>
<td><strong>0.2418</strong></td>
</tr>
<tr>
<td><strong>TP3,AP2</strong></td>
<td><strong>0.0694</strong></td>
<td><strong>0.1224</strong></td>
<td><strong>0.1238</strong></td>
<td><strong>0.2115</strong></td>
<td><strong>0.1932</strong></td>
<td><strong>0.2444</strong></td>
</tr>
<tr>
<td><strong>TP3,AP3</strong></td>
<td><strong>0.1133</strong></td>
<td><strong>0.1909</strong></td>
<td><strong>0.1267</strong></td>
<td><strong>0.2156</strong></td>
<td><strong>0.2490</strong></td>
<td><strong>0.2879</strong></td>
</tr>
<tr>
<td><strong>TP3,AP4</strong></td>
<td><strong>0.0693</strong></td>
<td><strong>0.1218</strong></td>
<td><strong>0.1183</strong></td>
<td><strong>0.2109</strong></td>
<td><strong>0.1854</strong></td>
<td><strong>0.2430</strong></td>
</tr>
<tr>
<td><strong>TP3,AP5</strong></td>
<td><strong>0.0701</strong></td>
<td><strong>0.1258</strong></td>
<td><strong>0.1255</strong></td>
<td><strong>0.2188</strong></td>
<td><strong>0.1955</strong></td>
<td><strong>0.2524</strong></td>
</tr>
<tr>
<td>Baseline</td>
<td>0.3620</td>
<td>0.4416</td>
<td>0.3665</td>
<td>0.4472</td>
<td>0.7285</td>
<td>0.6285</td>
</tr>
</tbody>
</table>
Two hops results

Fig. Occurrence of diffm and diffc in two data sets using TP3 and AP2
Results analysis

- Overall, performance in the Epinions.com dataset is better than in the Twitter dataset.
  - Epinions.com has direct rating information.
  - Accuracy of sentiment analysis tool.
  - Same criteria for all the user’s tweets.

- Formulas perform differently in the two datasets, e.g. TP1 and AP4.
  - Different user trust behavior patterns in different datasets or applications.
Select recommendation from trustworthy friends

- Weighted mean for trust aggregation (two hops)
- Besides using trustworthiness as a selection criteria, we also take confidence into account.

\[ m^i_s = \frac{\sum_{m^i_j \geq \text{th} \& \ c^i_j \geq \text{th}} m^i_j \cdot m^j_i}{\sum_{m^i_j \geq \text{th} \& \ c^i_j \geq \text{th}} m^i_j} \]
Stock Market Twitter Application

- Stock market has very reliable and available data to be used as ground truth.
- In this case, decision making is investing in stock market.
- Twitter sentiment is widely used to analyze stock market in literature.
- So we apply our trust management framework to analyze stock market data.
  - From twitter interactions, we model trust among pairs.
  - From trust among pairs, we aggregate, using our trust management system, the global trust/influence of each user from the community.
  - This global trust means how good each user is to predict stock market, we use this as a weighting factor.
Twitter trust network

- Twitter users from three groups: StockTwits, FinancialTimes, and MarketWatch.
- Build user-to-user trust network.
- Calculate users' power $P_u = \sum_{ui \in IN_u \& m_{ui} \geq 0.5} m_{ui} \cdot c_{ui}$
Twitter sentiment valence

- Stock market is related with people’s public mood.
- In Twitter, we measure people’s mood or sentiment by positive and negative tweets.

- Differentiating users by their power/influence.

\[ TSV = \log\left(\frac{1 + P}{1 + N}\right) \]

\[ TSV = \log\left(\frac{1 + \sum_{p \in PS} \text{Power}(\text{up} | \text{up posts } p)}{1 + \sum_{n \in NS} \text{Power}(\text{un} | \text{un posts } n)}\right) \]
Financial data

- 8 firms: Apple Inc (AAPL), Amazon.com Inc (AMZN), Alphabet Inc Class C (GOOG), Facebook Inc (FB), Netflix Inc (NFLX), Gilead Sciences Inc (GILD), General Electric Corp (GE), and Microsoft Corp (MSFT)

- Price change:
  \[ R_d = \frac{Price_d - Price_{d-1}}{Price_{d-1}} \]

- Abnormal return:
  \[ AR_d = R_d - E[R_d] \]

- Expected return is estimated using market model.
  \[ E[R_d] = \alpha + \beta \times RSP_d \]
## Number of tweets per firm

<table>
<thead>
<tr>
<th>Firm</th>
<th>Total number of tweets</th>
<th>Average number of daily tweets</th>
<th>Maximum number of daily tweets</th>
<th>Minimum number of daily tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>61,807</td>
<td>370.1018</td>
<td>2,653</td>
<td>101</td>
</tr>
<tr>
<td>FB</td>
<td>24,047</td>
<td>143.9940</td>
<td>1,089</td>
<td>37</td>
</tr>
<tr>
<td>GOOG</td>
<td>19,461</td>
<td>116.5329</td>
<td>704</td>
<td>29</td>
</tr>
<tr>
<td>NFLX</td>
<td>15,964</td>
<td>95.5928</td>
<td>665</td>
<td>13</td>
</tr>
<tr>
<td>AMZN</td>
<td>13,943</td>
<td>83.4910</td>
<td>912</td>
<td>14</td>
</tr>
<tr>
<td>GE</td>
<td>9,091</td>
<td>54.4371</td>
<td>491</td>
<td>10</td>
</tr>
<tr>
<td>MSFT</td>
<td>8,087</td>
<td>48.4251</td>
<td>567</td>
<td>9</td>
</tr>
<tr>
<td>GILD</td>
<td>7,329</td>
<td>43.8862</td>
<td>483</td>
<td>4</td>
</tr>
</tbody>
</table>
# Pearson correlation

Table 3: Comparison of Pearson correlation coefficients for eight firms

<table>
<thead>
<tr>
<th>Firms</th>
<th>$TSV_{equal}$</th>
<th>$TSV_{followers}$</th>
<th>$TSV_{power}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCC</td>
<td>p-value</td>
<td>PCC</td>
</tr>
<tr>
<td>AAPL</td>
<td>0.3370</td>
<td>$8.4 \times 10^{-6}$</td>
<td>0.3969</td>
</tr>
<tr>
<td>FB</td>
<td>0.0662</td>
<td>0.395</td>
<td>0.0544</td>
</tr>
<tr>
<td>GOOG</td>
<td>0.1830</td>
<td>0.018</td>
<td>0.1295</td>
</tr>
<tr>
<td>NFLX</td>
<td>0.1416</td>
<td>0.068</td>
<td>0.1758</td>
</tr>
<tr>
<td>AMZN</td>
<td>0.1314</td>
<td>0.091</td>
<td>0.3949</td>
</tr>
<tr>
<td>GE</td>
<td>0.0401</td>
<td>0.610</td>
<td>0.0043</td>
</tr>
<tr>
<td>MSFT</td>
<td>0.0533</td>
<td>0.494</td>
<td>0.1035</td>
</tr>
<tr>
<td>GILD</td>
<td>0.0969</td>
<td>0.213</td>
<td>0.0305</td>
</tr>
</tbody>
</table>
Stock abnormal returns might exhibit auto-correlation property, we also construct a linear regression model which includes both Twitter sentiment valence and historical abnormal returns.

Use previous 3 days’ AR as control variable

\[ AR_d = \alpha + \beta \times TSV_d + \gamma \times CV + \varepsilon_d \]

\[ AR_d = \alpha + \beta \times TSV_d + \sum_{i=1}^{3} \gamma_i \times AR_{d-i} + \varepsilon_d \]
Figure 4.3. Pearson correlation between AMZN’s abnormal returns and trust network power based Twitter sentiment valence
Summary or Twitter Stock Market Application

- We verified the hypothesis that the users reputation, built by the inter trust among them, using our trust management system, helps in making better analysis of abnormal stock returns.

- Compared with treating all the authors equally or simply weighting authors by the number of their followers, we could see that our trust network based mechanism could amplify the correlation between a specific firm's Twitter sentiment valence and the firm's stock abnormal returns.

- Limitation: our study showed that when the number of tweets about a firm is very small, the Twitter sentiment valence might not be able to reflect the stock market.
Cloud platforms

- Multi-clouds environment, e.g. collaboration among different organizations.
- A single task might be distributed over multiple computing nodes. So, there exist flows among computing nodes.
- A single computer node may be shared by multiple tasks. So, multiple tasks share computing resources and might affect each other.
Flow anomaly detection

- We treat flows as atomic measurements.

- Anomalous flow detection:
  - Distance based approaches, such as pre-defined profile and outlier detection.
  - Machine learning approaches, such as classification and clustering.
  - Output can be either continuous anomaly score or discrete labels, i.e. normal vs. abnormal.
Flow-level trust

- Given a set of anomaly detection results $M = \{m_1, m_2, ..., m_k\}$
- Also, we divide them by time windows.
  - When there is a new time window generated, previous time windows will be discounted by a forgetting factor $\sigma$, and $\sigma \leq 1$.
- Trustworthiness is the weighted mean of measurement results, where weights are from forgetting factors.

$$m = \frac{\sum_{i=1}^{k} w_i \cdot m_i}{\sum_{i=1}^{k} w_i} \quad r = \sqrt{\frac{\sum_{i=1}^{k} w_i \cdot (m_i - m)^2}{(\sum_{i=1}^{k} w_i) \cdot (\sum_{i=1}^{k} w_i - 1)}}$$
Node-level trust

- Trust of nodes is determined by both their incoming/outgoing flows and the tasks that are running on them.
  - If a node sends/receives a large amount of anomalous flows, it may execute some malicious missions or it may be compromised.
  - If malicious tasks are running on the node, it might be compromised later even though it has not exhibit malicious behaviors.
- For a node, given a set of running tasks $\{task_1, task_2, ... task_n\}$ and incoming flows $flow_1$ and outgoing flows $flow_0$.
- We treat all the incoming and outgoing flows as measurements to assess flow-level trust for nodes.
- Trust of nodes incorporate both flow-level trust and tasks’ trust.

$$m_{node} = \frac{w_{flow} * m_{flow} + \sum_{i=1}^{n} w_{task} * m_{task_i}}{w_{flow} + \sum_{i=1}^{n} w_{task}}$$

$$c_{node} = 1 - 2 * \sqrt{\left(\frac{w_{flow} * (1 - c_{flow})}{2 * (w_{flow} + \sum_{i=1}^{n} w_{task})}\right)^2 + \sum_{i=1}^{n} \left(\frac{w_{task} * (1 - c_{task_i})}{2 * (w_{flow} + \sum_{i=1}^{n} w_{task})}\right)^2}$$
Task-level trust

- Trust of tasks is determined by both their incoming/outgoing flows and the nodes that they are running on.
  - If a task sends/receives a large amount of anomalous flows, it may be a malicious mission.
  - If the nodes that the task is running on are compromised, it might be affected by the nodes.
- For a task, given a set of nodes $\text{Node} = \{\text{node}_1, \text{node}_2, ... \text{node}_N\}$, and incoming flows $\text{flow}_I$ and outgoing flows $\text{flow}_O$.
- We treat all the incoming and outgoing flows as measurements to assess flow-level trust for tasks.
- Trust of tasks incorporate both flow-level trust and nodes' trust.

$$m_{\text{task}} = \frac{w_{\text{flow}} * m_{\text{flow}} + \sum_{i=1}^{N} w_{\text{node}} * m_{\text{node}_i}}{w_{\text{flow}} + \sum_{i=1}^{N} w_{\text{node}}}$$

$$c_{\text{node}} = 1 - 2 \sqrt{\left(\frac{w_{\text{flow}} * (1 - c_{\text{flow}})}{2 * (w_{\text{flow}} + \sum_{i=1}^{n} w_{\text{node}})}\right)^2 + \sum_{i=1}^{n} \left(\frac{w_{\text{node}} * (1 - c_{\text{node}_i})}{2 * (w_{\text{flow}} + \sum_{i=1}^{n} w_{\text{node}})}\right)^2}$$
An attack example

- For each link, we assume that it has 10 flows.
- We let the forgetting factor $\sigma = 0.8$, and $w_{flow} = 2 \cdot w_{task}$

$$w_{flow} = 2 \cdot w_{node}$$
TRUST INFORMATION FOR ALL THE NODES AND TASKS IN 4 TIME WINDOWS

<table>
<thead>
<tr>
<th></th>
<th>TW0(m, c)</th>
<th>TW1(m, c)</th>
<th>TW2(m, c)</th>
<th>TW3(m, c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(0.84, 0.92)</td>
</tr>
<tr>
<td>T2</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(0.76, 0.90)</td>
<td>(0.57, 0.91)</td>
</tr>
<tr>
<td>T3</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(0.84, 0.92)</td>
</tr>
<tr>
<td>T4</td>
<td>(1,1)</td>
<td>(0.86, 0.93)</td>
<td>(0.65, 0.93)</td>
<td>(0.50, 0.95)</td>
</tr>
<tr>
<td>T5</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(0.72, 0.88)</td>
<td>(0.60, 0.91)</td>
</tr>
<tr>
<td>T6</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(0.79, 0.91)</td>
</tr>
<tr>
<td>N1</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(0.84, 0.92)</td>
<td>(0.71, 0.92)</td>
</tr>
<tr>
<td>N2</td>
<td>(1,1)</td>
<td>(0.87, 0.93)</td>
<td>(0.65, 0.93)</td>
<td>(0.47, 0.94)</td>
</tr>
<tr>
<td>N3</td>
<td>(1,1)</td>
<td>(0.72, 0.88)</td>
<td>(0.60, 0.91)</td>
<td>(0.43, 0.92)</td>
</tr>
<tr>
<td>N4</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(0.77, 0.88)</td>
</tr>
<tr>
<td>N5</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(0.85, 0.93)</td>
<td>(0.65, 0.93)</td>
</tr>
</tbody>
</table>
Trustability - Trust by Redundancy

- Redundancy is a basic requirement in many networking frameworks in order to provide reliable services.

- Trustability - guaranteeing a level of trust in face of attacks

\[ P(EA \cup EB) = P(EA) + P(EB) - P(EA \cap EB) \]

\[ m = m_1 + m_2 - m_1 \times m_2 \]

\[ c = 1 - 2 \times \sqrt{(\frac{(1 - m_2) \times (1 - c_1)}{2})^2 + \left(\frac{(1 - m_1) \times (1 - c_2)}{2}\right)^2} \]
Trustability assessment

To consider m and c together.

Algorithm 1: Trust-Reliability Assessment Algorithm

\[
\text{Input: } m; c; \text{lamda}_1; \text{lamda}_2; m\text{threshold}; \\
\text{Output: Trust-Reliability} \\
\]

1. if \( m \geq m\text{threshold} \):
2. then
3. \hspace{1em} \text{lamda} = \text{lamda}_1 \\
4. end
5. else
6. \hspace{1em} \text{lamda} = \text{lamda}_2 \\
7. end
8. \text{Normalizedmc} = \frac{2 \times (m - 0.5) \times c + 1}{2};
9. \text{Trust-Reliability} = \exp(- \text{lamda} \times (1 - \text{Normalizedmc}));
10. return Trust-Reliability;
Trustability Assessment

- Fixed $c$, trust-reliability assessment vs. $m$
Trustability Assessment

- Fixed m, trust-reliability assessment vs. c
Summary

- We provided a way for cloud vendors to estimate nodes and tasks' trust.

- Monitoring trust information can help to address attacks, e.g. migrating tasks from suspect nodes to trustworthy nodes.

- Tradeoff between trust-reliability and cost of resources can be used for dynamically allocating resources.
Making a decision regarding food-energy-water

- **Challenges:**
  - Complex decision (many parameters)
- **Multiple Stakeholders**
  - Difficult to optimize and make everyone happy
- **Sustainability**
  - Long- vs short-term optimizations
- **Human Factor**
  - Optimization equations may not be perfect
  - They may require (possibly long) time to mature
- **Negotiation among stakeholders**
  - Discussion
  - Finding out new factors/parameters
  - Agreement/consensus may result in stakeholder satisfaction at least at some degree
Building Trust Among Stakeholders

Solution Set

Ratings

Stakeholders/Actor

Actor

Decision

Trust

Actors
Collective Decision Making


- Adaptation planning problem for a community of large number of FEW stakeholders is examined in the regions of Umatilla and Morrow counties in Oregon.
A Solution

- A solution is, simply, an instance where parameters have values.
- Solutions are pre-optimized using environmental models.
- In this simulation, there are several parameters:
  - Ground water
  - Surface water
  - Crop choice
  - Fertilizer choice
- We experiment with 5 stakeholders, there will be a total of 20 parameters.
- Also, there are constraints on the total amount of ground/surface water.
Trust-Based Decision Support System

- In solution proposal stage, actors select a solution based on several criteria such as the profit and the environmental effects of the solutions.
- In solution rating stage, actors give the opinions on the proposed solutions as the ratings.
- In trust measurement stage, ratings are considered as the measurements, and the trust is calculated for each actor.
- Then, if there needs to be a next round, actors receive their updated trust and evaluate it which can have impact on their solution in the next round.
Trust Pressure Feedback Loop

- The process seen by an actor
- The decision making as a negotiation and trade-offs among actors
Example of a Solution Set

- A section of the sample solution set for an actor is given where each solution has a unique ID and differs in several parameters such as environmental protection values and profit which are normalized.

<table>
<thead>
<tr>
<th>Solution ID</th>
<th>Iteration</th>
<th>Water Amount</th>
<th>Environmental Protection</th>
<th>Profit</th>
<th>Solution Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>2628</td>
<td>4</td>
<td>396</td>
<td>0.745260</td>
<td>1.00000</td>
<td>1</td>
</tr>
<tr>
<td>2876</td>
<td>4</td>
<td>396</td>
<td>0.024715</td>
<td>0.98131</td>
<td>0</td>
</tr>
<tr>
<td>1884</td>
<td>3</td>
<td>375</td>
<td>0.762260</td>
<td>0.91892</td>
<td>2</td>
</tr>
<tr>
<td>3372</td>
<td>5</td>
<td>355</td>
<td>0.776400</td>
<td>0.87439</td>
<td>3</td>
</tr>
<tr>
<td>2132</td>
<td>3</td>
<td>375</td>
<td>0.914901</td>
<td>0.86310</td>
<td>0</td>
</tr>
<tr>
<td>3124</td>
<td>5</td>
<td>355</td>
<td>0.993470</td>
<td>0.85475</td>
<td>4</td>
</tr>
<tr>
<td>1636</td>
<td>3</td>
<td>375</td>
<td>0.995860</td>
<td>0.85414</td>
<td>5</td>
</tr>
<tr>
<td>7588</td>
<td>11</td>
<td>339</td>
<td>0.994820</td>
<td>0.85340</td>
<td>0</td>
</tr>
</tbody>
</table>
Distance and Power Metrics

- We introduce 2 metrics which are the distance between solutions and the power of a solution.
  - Distance metric (Dist) is calculated considering the Euclidean distance between the selected parameters of the solutions such as environmental protection value (e).
    \[ \text{Dist}(d, c) = \sqrt{\sum_{i}^{n} (e^d_i - e^c_i)^2} \]
  - Power of a solution (Pow) is calculated as the weighted average of the ratings where the weights are the trust of the proposers and the proposing is considered as giving a rating of 1.
    \[ \text{Pow}(s) = \frac{\sum_{i}^{P} m_i + \sum_{i}^{R} m_ir_s}{\sum_{i}^{P} m_i + \sum_{i}^{R} m_i} \]
Scenario for Significance of Trust Sensitivity

- Trust sensitivity is defined as the willingness of the actors to increase their trust to their desired level of trust.

- For example, if two actors currently have the same level of trust, the one with high trust sensitivity will try to increase her trust more than the other one by proposing a solution that would increase her trust more than the other actor.

- We designed the scenario for different trust levels from 0 to 1 with 0.1 increments.

- Also, we showed the effect of the round limitations of the decision making on the distance metric.
Scenario for Significance of Trust Sensitivity

As the trust sensitivity increases, the distance of the proposed solutions to the desired level decreases which means actors propose solutions that are better in terms of the given criteria.

The reason behind the curves to divert from having a smooth decrease is the discrete nature of the solutions.

Also, 5 rounds can be considered as the sensible number of rounds for a FEW decision making where 10 rounds might be a decision making with additional rounds due to remaining conflicts.
Social Networks
Use Trust for Filtering Attackers

Fig. 3. Overall damage for different confidence among attackers and other nodes in the case of simple attack
Use Machine Learning to Cluster Social Networks

- To divide the social network graph into clusters, spectral clustering method is used. This clustering method has very good performance in graph-based clustering and is a combination of k-means clustering and DBSCAN.
- After these clusters are detected, they are analyzed based on trust mechanism and other features to label them as the clusters of good users or the cluster of fake users.
Metrics

- To develop the evaluation metrics, we have assumed that fake users tend to establish more links between themselves and very few with real users and the trust propagation of these connections are extremely different from the connections of real users in social network.
- Based on these assumptions following are the characteristics which we have used to evaluate the clusters:
  - Density of the Clusters
  - Cluster formation based on time
  - Variation of trust over time
Results of the analysis shows that the trust and density of Cluster 3 is very high as compared to cluster 1 and cluster 2, so the probability of users in cluster 3 being fake is high.

To further validate our analysis we plot the graph of trust values of the cluster over time, to see the variation of trust which is shown in figure 3.
Fig 3: Change of Trust Values over time. Clusters with high trust value during creation and less fluctuation over time are more probable of being clusters of fake users.
Trust-Based Human-Machine Collaboration Mechanism for Predicting Crimes

1. The first step is to forecast crime hotspots using algorithms.
2. The second step is to take police feedback/input on the predictions made by the machine.
3. Third step is to evaluate both police and machine predictions based on actual crime data and calculate trust.
Crime Prediction System using Police in Loop

- We have used modulated Hawkes process to model crime hotspots which is a mix of multivariate model and point process model.
- This model uses spatial covariates like geographical data, census variables to link with the risk factor of different crime types.
- This model outputs the conditional intensity of different crime types that can happen in the area.

\[
\lambda_{g,c}(t) = \mu_{g,c} + \sum_{t > t_i} \theta_c w_c \exp(-w_c(t - t_i))
\]

- $\lambda_{g,c}(t)$ is conditional intensity
- $c_i$ is crime type
- $g$ is the grid cell
- $t_i$ is time
- $x_i$ is the spatial location
- $\mu_{g,c}$ is background intensity

Arjan Durresi

2- Trustworthy AI- 66

Trusted CI Webinar 4/27/2020
Police Feedback

- The output from this crime prediction system is then shown to the police to take their feedback about the predictions made by the machine model.
- Police will see these predictions and use their experience and ground knowledge to act towards these predictions.
- Police will either rely on the prediction model fully or not trust these predictions.
Decision Making Based on Trust

- We have used evaluation metrics that can be used to measure the predictive accuracy of both machine and the police.

- In our case, we have used Recall and Precision metrics to capture undetected crimes (false negatives) and false alarms (false positives).

\[
\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}
\]

\[
\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}
\]

\[
F1\text{Score} = \frac{2 \times (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}}
\]
Scenario: Police performed better than Machine

- In some regions where police have performed better than the machine. This is because police cognitive ability has been able to minimize the false positives i.e. false alarms generated by the machine model and still able to predict all the crimes.

- Trust information will help the police not to overtrust machine predictions as they are not very reliable in this region.

It shows the F1 score of police is higher than the machine because police can decrease false positives generated by the machine.

It shows that the trust of police is higher than the machine. This is because for this region police have better insight for crimes, than the machine prediction model.
Summary

- Trustworthy Decision Making and AI
- Trust can be a good platform for collaboration among Humans and Machines
- Use in cloud
- Water - Food Energy Decision Making
- Filtering fake users in social networks
- Predicting crime
Questions?

Please take our survey
Conference Updates

- GPN Annual Meeting
  May 19-21 in Kansas City, MO - **Online**
- PEARC20
  July 26-30 in Portland, OR - **Online**
- Trusted CI NSF Summit
  September 22-24 in Bloomington, IN - No updates yet
About the Trusted CI Webinar series

To view presentations, join the announcements mailing list, or submit requests to present, visit: https://trustedci.org/webinars

The next webinar is May 18th at 11am Eastern (1 week early)
Topic: Software Assurance
Speaker: Barton Miller