Big Data, Big Brother, and Systemic Risk Measurement and Management

Andrew W. Lo, MIT

Yale CPSC 458a Guest Lecture
Technology and the Financial System

Buttonwood Agreement
May 17, 1792
Technology and the Financial System

Wall Street Before the Invention of the Telegraph

Wall Street After the Invention of the Telegraph

Source: Leinweber, 2009, *Nerds on Wall Street*. 
Technology and the Financial System

New Challenges to Financial Regulation

- Capital requirements are harder to implement
- New risks to financial stability have been created
- Fairness and privacy issues have emerged
- Speed of financial innovation has increased; speed of regulatory innovation has not kept pace, e.g., HFT
- We need better “regulatory technology”: (1) risk transparency; (2) adaptive regulations; (3) framework for financial regulation (“regulation science”)

⇒ Financial System 2.0
The Need for Big Data
“Nobody Knows Anything”

The May 6, 2010 “Flash Crash”
“Nobody Knows Anything”

Accenture plc, Market Depth, Aggressive Buys, and Price

2:40pm - 2:55pm

© 2015 by Andrew W. Lo
All Rights Reserved
“Nobody Knows Anything”

September 30, 2010

FINISH REGARDE THE MARKET EVE OF MAY 6, 2010

Waddell ATT TIMS

April 3, 2013 5:43 pm

Flash crash exp By Philip Stafford


April 21, 2015

THE WALL STREET JOURNAL.

MARKETS

The ‘Flash Crash’ Charges Filed

Authorities say a trader in the U.K. helped trigger wild swings that shook markets in May 2010

"Nobody Knows Anything"

On Oct. 15, the yield on the 10-year Treasury note tumbled to its biggest one-day decline since 2009.

Wild Ride
Investors and regulators are trying to identify the reasons behind a plunge in Treasury yields.

Trading volumes in Treasury futures surged that day...
Volume

...raising questions about the role of high-speed traders.
Electronic trading of Treasurys as a share of total trading volumes

Sources: Tradeweb (intraday yields); CME Group (futures volume); TABB Group (trading share)
The Promise of Big Data
Big Data for Consumer Credit

- $3.4T of consumer credit outstanding as of Mar 2015
- $889B of revolving consumer credit outstanding as of Mar 2015
- 38% of households carry positive credit card balance in 2013 ($5,700)
Big Data for Consumer Credit

Standard Credit Scores Are Too Insensitive
Big Data for Consumer Credit

Anonymized Data from Large U.S. Commercial Bank

Transaction Data

<table>
<thead>
<tr>
<th>By Category</th>
<th>By Category</th>
<th>By Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortgage payment</td>
<td>Hotel expenses</td>
<td>Bar Expenses</td>
</tr>
<tr>
<td>Credit card payment</td>
<td>Travel expenses</td>
<td>Fast Food Expenses</td>
</tr>
<tr>
<td>Auto loan payment</td>
<td>Recreation (golf)</td>
<td>Total Rest/Bars/Fast-Food</td>
</tr>
<tr>
<td>Student loan payment</td>
<td>Department Stores Expenses</td>
<td>Healthcare related expenses</td>
</tr>
<tr>
<td>All other types of loan payment</td>
<td>Retail Stores Expenses</td>
<td>Gas stations expenses</td>
</tr>
<tr>
<td>Other line of credit payments</td>
<td>Clothing expenses</td>
<td>Vehicle expenses</td>
</tr>
<tr>
<td>Brokerage net flow</td>
<td>Discount Store Expenses</td>
<td>Car and other insurance</td>
</tr>
<tr>
<td>Dividends net flow</td>
<td>Big Box Store Expenses</td>
<td>Drug store expenses</td>
</tr>
<tr>
<td>Utilities Payments</td>
<td>Education Expenses</td>
<td>Government</td>
</tr>
<tr>
<td>TV</td>
<td>Total Food Expenses</td>
<td>Treasury (e.g. tax refunds)</td>
</tr>
<tr>
<td>Phone</td>
<td>Grocery Expenses</td>
<td>Pension Inflow</td>
</tr>
<tr>
<td>Internet</td>
<td>Restaurant Expenses</td>
<td>Collection Agencies</td>
</tr>
<tr>
<td>Collection Agencies</td>
<td>Unemployment Inflow</td>
<td></td>
</tr>
</tbody>
</table>

Balance Data

Checking Account Balance
Brokerage Account Balance
Saving Account Balance
CD Account Balance
IRA Account Balance

Credit Bureau Data

File Age
Credit Score
Open/Closed Flag & Date of Closure
Bankruptcy (Date & Code)
MSA & Zip

Type (CC, MTG, AUT, etc)
Age of Account
Balance
Limit if applicable
Payment Status
48-Month Payment Status History

1% Sample = 10 Tb!
Big Data for Consumer Credit

Graph showing the probability of 90+ delinquency in subsequent 6 months, with two lines: unconditional and conditional on income drop.
Big Data for Consumer Credit

- Decision tree
- Logistic regression
  - Outputs a probability of delinquency
- Random forest
- Clustering/segmentation
  - Can be combined with other models

- Software:
  - LIBLINEAR (National Taiwan University) [http://www.csie.ntu.edu.tw/~cjlin/liblinear/](http://www.csie.ntu.edu.tw/~cjlin/liblinear/)
Big Data for Consumer Credit

**Inputs describing consumer $j$**
- Consumer level categorical expenditures
- Consumer credit history & financial behaviors

**Forecast Model**

$F^*(x)$

**Credit Risk Forecast of consumer $j$**

$P(X_j)$: probability of consumer $j$ becoming 90+ days delinquent within next 3 months

---

12 Oct 2015

© 2015 by Andrew W. Lo
All Rights Reserved
Big Data for Consumer Credit

- Khandani, Kim, and Lo (2010)
- 600,000 credit cards per month; 40-hour runtime
Credit Forecasts Over Time

(a) Time series of actual and predicted 90-days-or-more delinquency rates (6-month)

(b) Time series of actual and predicted 90-days-or-more delinquency rates (12-month)
Current Research

Risk and Risk Management in the Credit Card Industry

Florentin Butaru2, Qingqing Chen2, Brian Clark5, Sanmay Das3, Andrew W. Lo3, Akhtar Siddique2

Abstract

Using account level credit-card data from eight major commercial banks over the period January 2009 to December 2012, we build decision tree models that use combined consumer tradeline, credit-bureau, and macroeconomic data to predict delinquency. We use our models to analyze and compare risk management practices and the drivers of delinquency across the banks, and find substantial heterogeneity in risk factors and sensitivities, implying that no single model applies across all eight institutions. Predictability of delinquency varies across institutions as does the efficacy of risk management, where efficacy is measured by the percentage of bad accounts that banks manage effectively. It appears that certain firms tend to be significantly more active and effective at managing the exposure of their credit-card portfolios. Banks are also quite different in their macroeconomic sensitivities. These differences are a function of their portfolio characteristics as well as risk management practices. These results suggest that supervision that is specific and individualized to an institution is likely to be more effective.

1Disclaimer: The statements made and views expressed herein are solely those of the authors and do not necessarily represent official policies, statements, or views of the Office of the Comptroller of the Currency, Washington University, MIT, RPI, AlphaSimplex, or their employees.
2U.S. Department of the Treasury, Office of the Comptroller of the Currency, Risk Analysis Division.
3Washington University, Department of Computer Science.
4Massachusetts Institute of Technology, Sloan School of Management and Computer Science and Artificial Intelligence Laboratory; Director, MIT Laboratory for Financial Engineering; AlphaSimplex Group, LLC.
5Rensselaer Polytechnic Institute (RPI), Lally School of Management

- 6 largest banks from Jan 2008 to Dec 2014
- Macro and institution-specific factors (137)
- 25 Tb of data
- Used to gauge quality of risk management across institutions
- Models vary greatly
- But you need **data!**
Journal of Business & Technology Law

Volume 10 | Issue 2

Law is Code: A Software Engineering Approach to Analyzing the United States Code

William Li
Pablo Azar
David Larochelle
Phil Hill
Andrew W. Lo
## Law As Code

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>United States Code</th>
<th>Software Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>Titles, Chapters, Sections, Sub-sections</td>
<td>Libraries, Classes, Sub-classes, Functions</td>
</tr>
<tr>
<td>Execution</td>
<td>Businesses, regulators, police, courts, interpret and execute laws</td>
<td>Computers interpret and execute software</td>
</tr>
<tr>
<td>Evolution</td>
<td>Congress passes laws that edit the U.S. Code</td>
<td>Programmers edit the software system’s codebase</td>
</tr>
<tr>
<td>Language</td>
<td>English (specialized “legal language”)</td>
<td>Computer-interpretable language (e.g. C, Java)</td>
</tr>
</tbody>
</table>
Law As Code

- “All general and permanent laws…”
- HeineOnline, U.S. Code 1925–2006 (does not include all regulations, CFR); Office of the Law Revision Counsel
- U.S. Code size:
  - 1925–2006: 36 gigabytes
  - For 2014 (from U.S. OLRC): 1.8 million sentences (41.4 million words, >200,000 pages)
- Comparison with other software:
  - Mozilla Firefox: 14.4 million lines of code
  - Linux kernel: 23.5 million lines of code
<table>
<thead>
<tr>
<th>Title Subject</th>
<th>Title Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 GENERAL PROVISIONS</td>
<td>26 INTERNAL REVENUE CODE</td>
</tr>
<tr>
<td>2 THE CONGRESS</td>
<td>27 INTOXICATING LIQUORS</td>
</tr>
<tr>
<td>3 THE PRESIDENT</td>
<td>28 JUDICIARY AND JUDICIAL PROCEDURE</td>
</tr>
<tr>
<td>4 FLAG AND SEAL, SEAT OF GOVERNMENT, AND THE STATES</td>
<td>29 LABOR</td>
</tr>
<tr>
<td>5 GOVERNMENT ORGANIZATION AND EMPLOYEES</td>
<td>30 MINERAL LANDS AND MINING</td>
</tr>
<tr>
<td>6 DOMESTIC SECURITY</td>
<td>31 MONEY AND FINANCE</td>
</tr>
<tr>
<td>7 AGRICULTURE</td>
<td>32 NATIONAL GUARD</td>
</tr>
<tr>
<td>8 ALIENS AND NATIONALITY</td>
<td>33 NAVIGATION AND NAVIGABLE WATERS</td>
</tr>
<tr>
<td>9 ARBITRATION</td>
<td>34 PATENTS</td>
</tr>
<tr>
<td>10 ARMED FORCES</td>
<td>35 PATRIOTIC AND NATIONAL OBSERVANCES, CEREMONIES, AND ORGANIZATIONS</td>
</tr>
<tr>
<td>11 BANKRUPTCY</td>
<td>36 PAY AND ALLOWANCES OF THE UNIFORMED SERVICES</td>
</tr>
<tr>
<td>12 BANKS AND BANKING</td>
<td>37 VETERANS BENEFITS</td>
</tr>
<tr>
<td>13 CENSUS</td>
<td>38 POSTAL SERVICE</td>
</tr>
<tr>
<td>14 COAST GUARD</td>
<td>39 PUBLIC BUILDINGS, PROPERTY, AND WORKS</td>
</tr>
<tr>
<td>15 COMMERCE AND TRADE</td>
<td>40 PUBLIC CONTRACTS</td>
</tr>
<tr>
<td>16 CONSERVATION</td>
<td>41 THE PUBLIC HEALTH AND WELFARE</td>
</tr>
<tr>
<td>17 COPYRIGHTS</td>
<td>42 PUBLIC PRINTING AND DOCUMENTS</td>
</tr>
<tr>
<td>18 CRIME AND CRIMINAL PROCEDURE</td>
<td>43 PUBLIC LANDS</td>
</tr>
<tr>
<td>19 CUSTOMS DUTIES</td>
<td>44 RAILROADS</td>
</tr>
<tr>
<td>20 EDUCATION</td>
<td>45 SHIPPING</td>
</tr>
<tr>
<td>21 FOOD AND DRUGS</td>
<td>46 TELEGRAPHS, TELEPHONES, AND RADIOTELEGRAPHS</td>
</tr>
<tr>
<td>22 FOREIGN RELATIONS AND INTERCOURSE</td>
<td>47 TERRITORIES AND INSULAR POSSESSIONS</td>
</tr>
<tr>
<td>23 HIGHWAYS</td>
<td>48 TRANSPORTATION</td>
</tr>
<tr>
<td>24 HOSPITALS AND ASYLUMS</td>
<td>49 NATIONAL AND COMMERCIAL SPACE PROGRAMS</td>
</tr>
<tr>
<td>25 INDIANS</td>
<td>50 WAR AND NATIONAL DEFENSE</td>
</tr>
<tr>
<td></td>
<td>51 NATIONAL AND COMMERCIAL SPACE PROGRAMS</td>
</tr>
</tbody>
</table>
Law As Code

Change in Size of U.S. Legal Code, by Title

U.S. Core Title

Year

count by year


count by year


© 2015 by Andrew W. Lo
All Rights Reserved
## The Five C’s of Software Development

<table>
<thead>
<tr>
<th>Principle</th>
<th>Proposed Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conciseness</strong></td>
<td>Number of words</td>
</tr>
<tr>
<td><strong>Cohesion</strong></td>
<td>Language perplexity</td>
</tr>
<tr>
<td><strong>Change</strong></td>
<td>Number of sections/subsections affected</td>
</tr>
<tr>
<td><strong>Coupling</strong></td>
<td>Size of cross-reference network core versus periphery</td>
</tr>
<tr>
<td><strong>Complexity</strong></td>
<td>Number of condition statements in code (McCabe’s complexity)</td>
</tr>
</tbody>
</table>
Law As Code

4. **Coupling:**

**Principle:** Modular code is more robust and easier to maintain than code with unnecessary cross-dependencies (reduces complexity)

**Metric:** Cross-reference network analysis, e.g., Baldwin, MacCormack, and Rusnak, (2013)

- Map the network of interdependencies of subcomponents of large software systems
- Develop measures of complexity, robustness, etc.
- Begin with notion of “strongly connected” nodes
Law As Code

- U.S. Code network has 36,670 sections (nodes)
- Core is 6,947 sections (18.9%)
# Law As Code

## Laws Passed in 111\textsuperscript{th} Congress with Highest Coupling

<table>
<thead>
<tr>
<th>Public Law Number</th>
<th>Title of Act</th>
<th>Total Number of Affected Core Sections</th>
</tr>
</thead>
<tbody>
<tr>
<td>111-5</td>
<td>American Recovery and Reinvestment Act</td>
<td>293</td>
</tr>
<tr>
<td>111-148</td>
<td>Patient Protection and Affordable Care Act</td>
<td>251</td>
</tr>
<tr>
<td>111-203</td>
<td>Dodd-Frank Wall Street Reform and Consumer Protection Act</td>
<td>232</td>
</tr>
<tr>
<td>111-312</td>
<td>Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act</td>
<td>199</td>
</tr>
</tbody>
</table>
Omnibus Appropriations Act (111–8)
Law As Code

Title 35 (Patents)
Law As Code

Title 26 (Internal Revenue Code)
Law As Code

Quality Assurance?? For Linux Kernel “Patches”:

- Designated individuals (“Maintainers”) oversee different versions
- Benevolent dictatorship: “Linus Torvalds is the final arbiter of all changes to the Linux kernel.” – www.kernel.org/doc/Documentation/SubmittingPatches
- What’s the equivalent for financial regulation?
The Threat of Big Data
Privacy vs. Transparency

Predicting Social Security numbers from public data

Alessandro Acquisti¹ and Ralph Gross
Carnegie Mellon University, Pittsburgh, PA 15213

Communicated by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, May 5, 2009 (received for review January 18, 2009)

Information about an individual’s place and date of birth can be exploited to predict his or her Social Security number (SSN). Using only publicly available information, we observed a correlation between individuals’ SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. The inferences are made possible by the public availability of the Social Security Administration’s Death Master File and the widespread accessibility of personal information from multiple sources, such as data brokers or profiles on social networking sites. Our results highlight the unexpected privacy consequences of the complex interactions among multiple data sources in modern information economies and quantify privacy risks associated with information revelation in public forums.

identity theft | online social networks | privacy | statistical reidentification

number (SN). The SSA openly provides information about the process through which ANs, GNs, and SNs are issued (1). ANs are currently assigned based on the zipcode of the mailing address provided in the SSN application form [RM00201.030] (1). Low-population states and certain U.S. possessions are allocated 1 AN each, whereas other states are allocated sets of ANs (for instance, an individual applying from a zipcode within New York state may be assigned any of 85 possible first 3 SSN digits). Within each SSA area, GNs are assigned in a precise but nonconsecutive order between 01 and 99 [RM00201.030] (1). Both the sets of ANs assigned to different states and the sequence of GNs are publicly available (see www.socialsecurity.gov/employer/stateweb.htm and www.ssa.gov/history/ssn/geocard.html). Finally, within each GN, SNs are assigned “consecutively from 0001 through 9999” (13) (see also [RM00201.030], ref. 1.)
Is There A Compromise Between Data Privacy and Transparency?
Secure Multi-Party Computation

\[
\begin{align*}
Y_1 &= S_1 + X_1 \\
Y_2 &= Y_1 + S_2 + X_2 \\
\vdots \\
Y_{n-1} &= Y_{n-2} + S_{n-1} + X_{n-1} \\
Y_n &= Y_{n-1} + S_n + X_n \\
\end{align*}
\]

Andrew

\[
\begin{align*}
Y_n &= Z_n \\
Z_1 &= Y_n - X_1 \\
Z_2 &= Z_1 - X_2 \\
\vdots \\
Z_{n-1} &= Z_{n-2} - X_{n-1} \\
Z_n &= Z_{n-1} - X_n \\
\end{align*}
\]

Alice

Bob

Andrew

Alice

Bob

\[
\frac{1}{n} \sum_{i=1}^{n} S_i = \frac{1}{n} Z_n
\]
Privacy and Transparency

Transparency and Privacy Can **Both** Be Achieved

- Individual data is kept private, e.g., RSA
- Encryption algorithms are “collusion-robust”
- Aggregate risk statistics can be computed using encrypted data
  - Means, variances, correlations, percentiles, Herfindahl indexes, VaR, CoVaR, MES, etc.
- Privacy is preserved, no need for raw data!
Privacy and Transparency

Real Estate Loans Outstanding

Graph showing the growth of individual bank values and aggregate values over time for different banks.
Privacy and Transparency

Real Estate Loans Outstanding

© 2015 by Andrew W. Lo
All Rights Reserved
Conclusion

- Technology has transformed everything!
- Financial markets are vastly better off
- But new challenges have emerged
- We can do better
- We have to do better
- Regulation has to account for technology and how it interacts with human behavior
- Computer science, law, and finance must collaborate to create the Financial System 2.0
Thank You!
Additional References


## Additional References