Research Challenges and Needed Resources for Data Mining in Financial Market Regulation

Henry Goldberg – Henry.Goldberg@FINRA.org
Special Projects Business Solutions Dept.

October 11, 2007 – NGDM ‘07

The Financial Markets Regulatory Perspective

- The Financial Industry Regulatory Authority (FINRA) is the largest non-governmental regulator for all securities firms doing business in the United States.
- FINRA regulates several Markets –Nasdaq, Amex, Over the Counter, Corporate and Municipal Bonds, Chicago Climate Exchange, dealing with Equities, Options, Bonds, other derivatives, and even CO2 emissions futures.
- In addition FINRA regulates all securities industry professionals doing business with the public in the US as well as all member firms of Nasdaq and NYSE.
- www.finra.org

Mining in an Ongoing Regulatory Environment

- **Break Detection**
  - Several systems provide regulatory analysts with a continual feed of alerts (or breaks) – detected episodes of interest, patterns of behavior, or outliers.
  - Breaks also provide a significant reduction of the data to the point where traditional data mining methods are tractable.
  - But they currently require manual encoding of knowledge, whether discovered or heuristic.
- **Predictive models to allow proactive regulation. Assessing risk rather than just detecting violations is especially important for fraud, manipulation, and focusing member regulation reviews.**
- **Presentation of discovered knowledge – methods must be able to support case presentation in court, relate discovered knowledge back to specific instances, examples, statistics to support validity.**

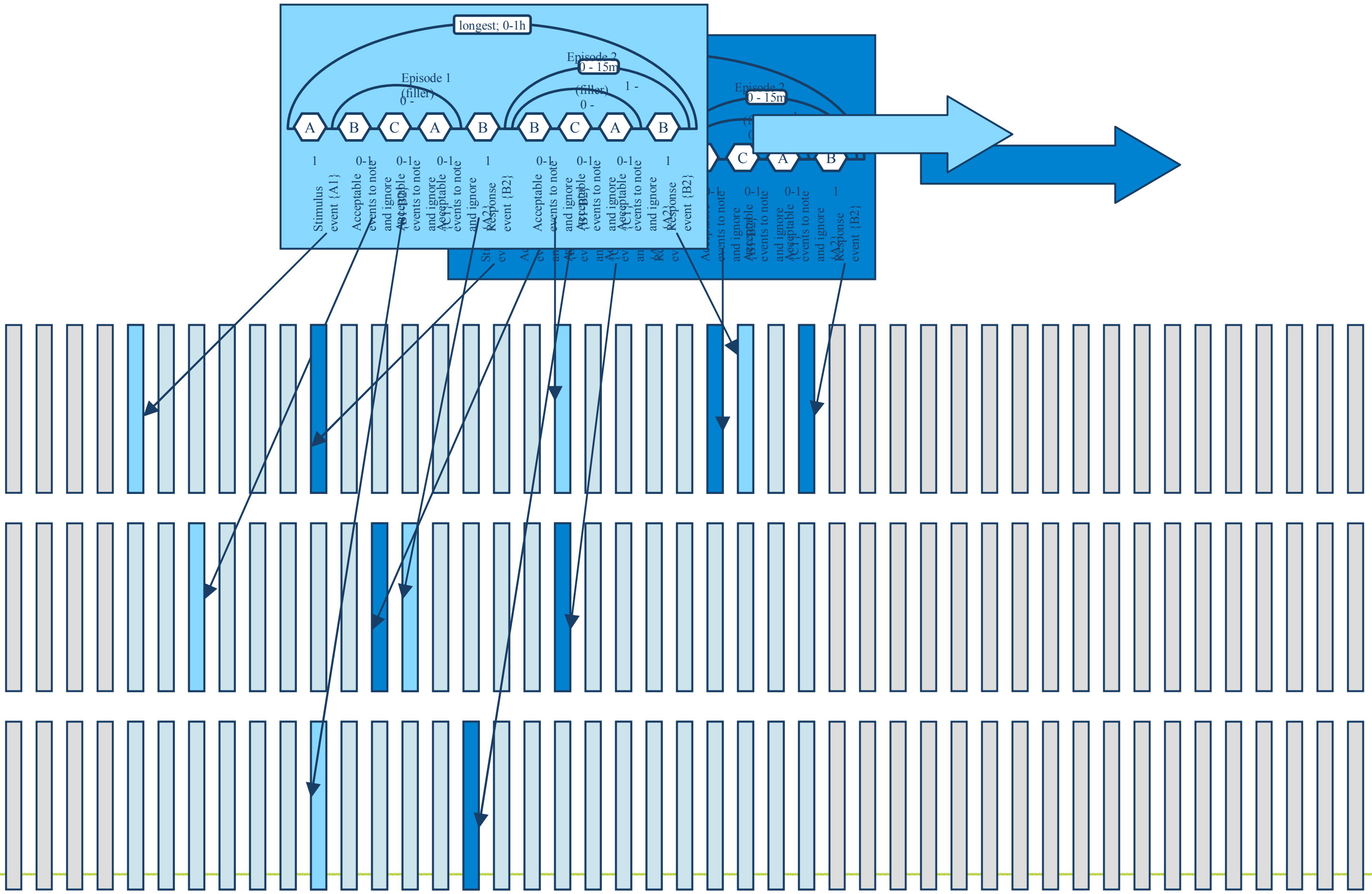
Challenges

- Data Volume – over 400 million transactions per trading day (trades, quotations, orders)
- Data Quality – in addition to the usual errors, there are data gaps due to market independence and fragmentation of order lifecycle.
- Data Linkage – data is often received asynchronously, especially order lifecycle events, requiring linkage.
- Data Volatility – Data ages quickly, some detection needs to be done in real time, some within a day or two.
- Domain Volatility – Constant and rapid change in infrastructure of the markets and participants, rules and interpretations, and participant behavior.
- Hidden Relationships – many scenarios of interest revolve around inferring the existence of relationships.

Hidden Relationships

- Routed Orders – order flow relationships
- Wash Trading – manipulation of prices and volume; money laundering
- Statistical Relational Models of brokerage firms and individuals (UMass/KDL)
- “Tribe analysis” – mining dynamic relationships
- Many detection tasks are really finding hidden relationships among scenarios

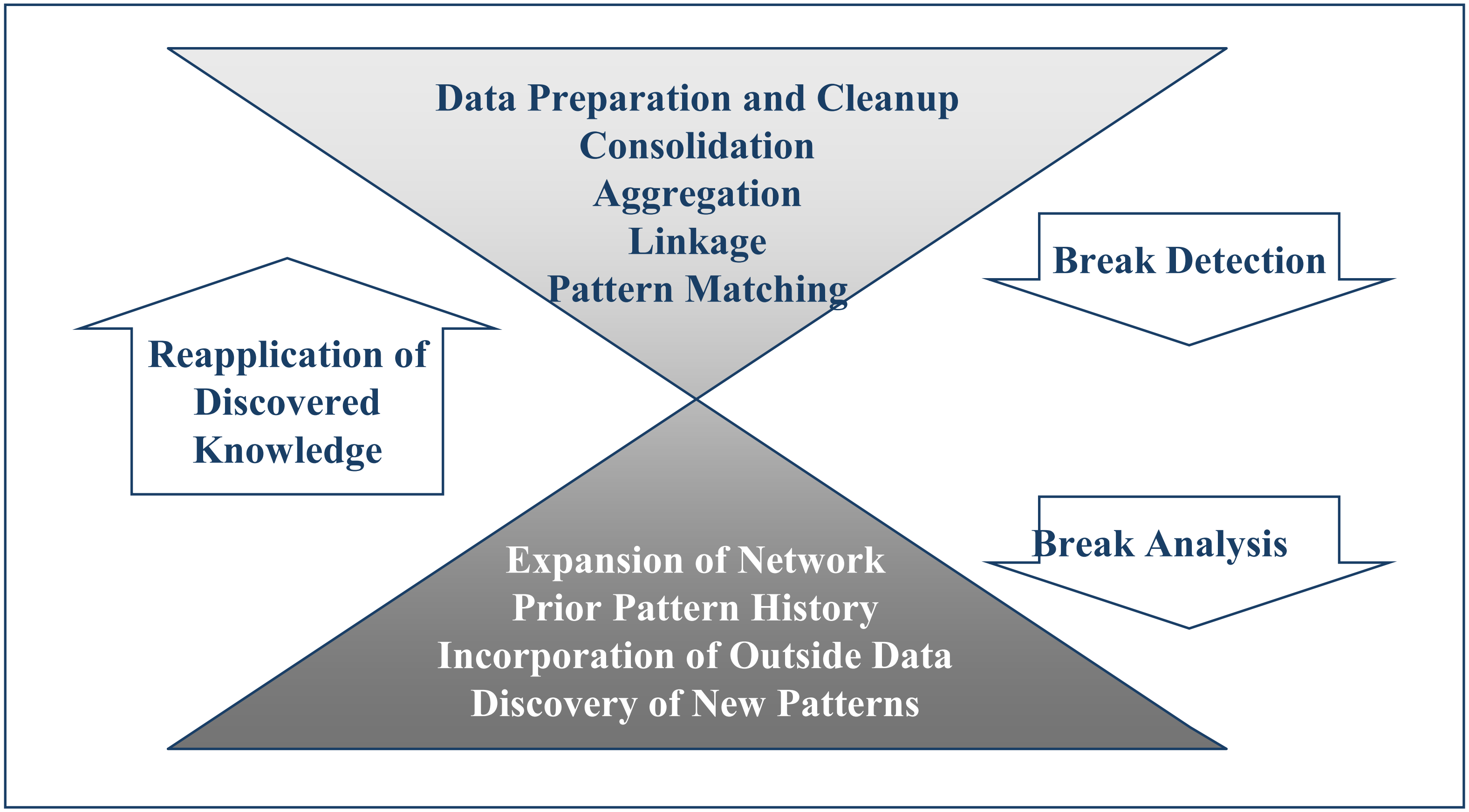
Relationships Among Transactions


longest; 0-1h
Episode 1
(filler)
0 -
Episode 2
0 - 15m
(filler)
0 -
1 -
A
B
C
A
B
B
C
A
B
1
0-1
0-1
0-1
1
0-1
0-1
0-1
1
Stimulus event {A1}
Acceptable events to note and ignore {B1}
Acceptable events to note and ignore {C1}
Acceptable events to note and ignore {A2}
Response event {B2}
Acceptable events to note and ignore {B2}
Acceptable events to note and ignore {C1}
Acceptable events to note and ignore {A2}
Response event {B2}
Episode 2
0 - 15m
C
A
B
0-1
0-1
0-1
1
Response event {B2}

Supporting the Regulatory Analyst

- Complex regulatory knowledge is expressed in the language of law that very often makes its translation to the technical system specifications difficult and error prone.
- (Semi-)automated support for analysis of alerts/cases; support for knowledge acquisition from expert analysts
- Systems for analyst training and advice
- We hope to develop a financial regulation ontology, and to begin applying it to alert analysis through apprentice learning and semantic analysis of analysts' decisions.
- "Formalization of Capital Markets Regulations Using Semantic and Rule-based Technologies" (Exprentis)

"We need to work on the second triangle"



Data Preparation and Cleanup
Consolidation
Aggregation
Linkage
Pattern Matching

Break Detection

Reapplication of Discovered Knowledge

Break Analysis

Expansion of Network
Prior Pattern History
Incorporation of Outside Data
Discovery of New Patterns

Conduct Rule 2440

In "over-the-counter" transactions, whether in "listed" or "unlisted" securities, if a member buys for his own account from his customer, or sells for his own account to his customer, he shall buy or sell at a price which is fair, taking into consideration all relevant circumstances, including market conditions with respect to such security at the time of the transaction, the expense involved, and the fact that he is entitled to a profit; and if he acts as agent for his customer in any such transaction, he shall not charge his customer more than a fair commission or service charge, taking into consideration all relevant circumstances, including market conditions with respect to such security at the time of the transaction, the expense of executing the order and the value of any service he may have rendered by reason of his experience in and knowledge of such security and the market therefor.

An Alert

On October 15, 2006, a brokerage firm, ABCD, purchased the XYZ bond in the market at a price of $70.40 from another broker-dealer, HIJK, (who in turn purchased the bond at a price of $69.625, charging ABCD a markup of just 1.11%) and subsequently sold the bond to a customer at a price of $73.40, resulting in a 4.26% markup. Both the buy and sell transactions had a reporting time of 11:04:00, suggesting a riskless principal transaction.

HIJK ← ???? @ 69.625
ABCD ← HIJK @ 70.40 (1.11%)
Customer ← ABCD @ 73.40 (overall 4.226%)

Analysis of the A[illegible]

ABCD stated there was “limited activity, with only a few trades in the market” for this security. However, the audit trail sys[illegible] four (34) [illegible] wo hundred sixty (260 [illegible] security in [illegible] proceeding week. Sixty-two (62) of these trades occ[illegible] on the same trade date as ABCD trades. This data indicates that the se[illegible]rity was liquid and readily available in the m[illegible] trade[illegible] ABCD markup (in relation to comparable “risk[illegible] r the highest during this period of time. The firm also indicated there was “volatility in the XYZ [illegible] additional pricing attention) however; the buy order for this transact[illegible] was time-stamped at 10:52:00 and HIJK was able to price the bond and purchase it from another broker-dealer within twelve (12) minutes (with an execution time of 11:04:00). Thus, the staff believes that this item should be considered a violation of Rule 2440 and IM-2440 and included in the further review and a potential regulatory action.

Retrieves contextual knowledge.

Concept of “liquidity”

Local comparison with similar entities

Local comparison with similar entities

Counter-evidence

Relates case to rule

We need Data Mining methods which…

- Operate continuously over dynamic data streams.
- Are not sensitive to gaps in data.
- Explain the source and support of the discovered knowledge.
- “Track” changes in the data – models continually adapt themselves as new data is analyzed.
- Find rare (rather than frequent) sub-sequences (-graphs, -sets).
- And support the Regulatory Analyst by explaining to and learning from her in her own language.

And Data Sharing Resources

- **Our most critical needs are time and brainpower. We look to the academic and research community to provide these. (We are hiring, by the way.)**
- **We've got data, which we can provide, but the demands of daily operations limit our ability to work with outside researchers to provision the data to their needs, explain the data model, and discuss the knowledge domains. (also $$$$)**
- **Perhaps we can build a "platform" for collaboration.**
  - Hardware, data transformation software, and knowledge engineering resources to manage the provision of real world, operational data to the research community.
  - Several commercial tool providers have "general" financial data models (proprietary).
  - Knowledge Management can provide a common language for regulatory domain knowledge.

Background

- [Friedland, Jensen] Finding tribes: Identifying close-knit individuals from employment patterns, *proc. KDD 2007*
- [Fast, etal.] Relational Data Pre-Processing Techniques for Improved Securities Fraud Detection, *proc. KDD 2007*
- [Senator, Goldberg] “Break Detection Systems.” *Handbook of Knowledge Discovery and Data Mining*, eds. Kloesgen & Zytkow, Oxford 2002.
- [Goldberg, etal.] “The NASD Securities Observation, News Analysis & Regulation System (SONAR),” *proc. IAAI 2003.*
- [Senator] “Ongoing Management and Application of Discovered Knowledge in a Large Regulatory Organization: A Case Study of the Use and Impact of NASD Regulation’s Advanced Detection System (ADS),” *proc. KDD 2000.*
- [Kirkland, etal.] “The NASD Regulation Advanced Detection System (ADS),” *proc. IAAI 1998.*