Anomaly Detection Technology Using BigGraph

● Bo Hu ● Aisha Naseer ● Takahide Matsutsuka

Many difficulties are encountered along all three axes of Big Data (volume, variety, and velocity), which limit the applicability of established technology. BigGraph is a research project and platform for realising the vision of an intelligent society, where requirements along all three axes can be accommodated. In the fraud detection scenario, working with Fujitsu UK & Ireland, BigGraph researchers consider different types of information, link external data sources as streams, and consolidate existing knowledge in order to identify undiscovered and fraudulent anomalies in Big Data.

1. Introduction

The three axes of Big Data, volume, variety, and velocity, have brought new dimensions to data exploitation and data analytics.[1)] New types of business intelligence taking full advantage of all three axes have emerged rapidly and have been used to address challenges that are deemed impossible and/or impractical when processing large amounts of data is not supported. BigGraph,[2)] a project initiated at the Fujitsu Laboratories of Europe Limited (FLE), provides software and expertise to customers, helping them to link together data silos in the enterprise architecture. A plethora of services can then be built around the well-connected enterprise data. As a data management platform underpinned by the large-scale Resource Description Framework (RDF),[3)] stream data analysis, and semantic technologies, BigGraph lowers the threshold of applying Linked Data principles over distributed data sources and thus unlocks hidden insights within an enterprise. Recent public sector interest in Linked Open Data (LOD) and other semantic technologies has been fuelled by national and international directives on sharing and reusing governmental data.[4)] In the context of fraud detection, BigGraph offers a new apparatus, one for detecting, from unprecedented perspectives, fraudulent social benefit claims.

Fraudulent claims account for a significant portion of claims received by private finance providers as well as public welfare authorities. It is estimated that fraud loss cost the UK public and private sectors 73 billion pounds in 2012, of which the loss due to fraudulent benefit and tax claims alone was estimated to be at least £1.6 billion. These amounts are expected to have increased significantly in 2013.[5)] Challenges of similar magnitude are observed in many other industrialised countries.[6),7)] Thus far, benefit fraud detection is largely performed within individual councils and is labour intensive. Though some information and communications technology (ICT) support is in place to relieve civil servants of manual data comparison and anomaly detection, given the scale and complexity of the problem, such support is very limited.

Current best practice is built on top of relational databases (RDBs) and structured query language (SQL) scripts. The relational database has recognised limitations as a solution basis for scenarios in which data is highly distributed and sizable and the model structures are evolving and de-centralised. New paradigms in data management, collected under the label "Big Data," offer a whole raft of technologies to interconnect heterogeneous data and reveal hidden links that are buried inside the seemingly tangled riddles of facts. This aligns perfectly with fraud detection wherein the challenge is to efficiently pinpoint small anomalies in Big Data, often on the basis of relationship patterns between data.

Here we report on the application of an anomaly detection technology using BigGraph in the public sector. Our main motivation is to demonstrate how BigGraph and Linked Data can be used to solve a typical analytical task in real-life settings, making it easier to detect fraud and anomalies.

2. Issues and challenges with anomaly detection

Several anomaly detection techniques have been developed and applied to various application domains.[8),9)] In recent years, fraudulent and false benefit claims have attracted more attention due to both political and technical issues.[10)] On the one hand, many developed countries have implemented firm austerity measures focused on cutting public spending, leading to rigid benefit claim evaluation and review. On the other hand, the push by the US and UK governments to publish government data online[11)] has fuelled better utilisation and transparency of public sector data, paving the way for innovative new fraud detection methodologies.

As a result, we are starting to witness the formation of strategic partnerships and alliances aimed at cross-sector and cross-council data sharing. The positive aspect of such a movement is that the integration of multiple data sources will enable the discovery of new fraud patterns. The downside is equally evident: current ICT supports are made obsolete and "smart" solutions must be implemented. Our analysis of technologies has highlighted five barriers and inefficiencies in current best practice.

1) Silo-ed information sources

Regardless of the efforts made in e-governance, progress in integrating public sector data is rather discouraging. Data warehousing is highly valued but not widely practiced due to security, privacy, and ownership concerns.

2) Model rigidity

With the type of fraudulent claims constantly evolving and the number of claims overwhelming the capacity of existing fraud detection tools, extensibility is always a challenge. This calls for flexibility in data schemata, anomaly patterns, and data sources. Newly confirmed fraudulent cases should be analysed, and their patterns should be easy to incorporate.

3) Data heterogeneity

Even within one local council, syntactical and semantic discrepancies are evident across different governmental data repositories that are normally created and curated by different departments and/or outsourced to different ICT vendors, with each having only a fragmented view of the domain of discourse.

4) Lack of specific analytic models

Most tools currently available were not designed and developed for social benefit fraud detection. They include generic on line analytical processing (OLAP) tools[12)] and business intelligence (BI) tools, which lack intuitive features for the social benefit fraud detection domain.

5) "Un-timeliness"

Current metadata repositories and RDBs are generally passive, offering data but no way to act on it. Defining and automatic invoking of complex business logic are supported by current technology, leading to a gap between detection and action. In many cases, this latency can have significant consequences in terms of monetary loss.

These issues and challenges have inspired us to conceive a new technology that enables organisations to dynamically describe unprecedented fraudulent cases and that facilitates integration, analysis, and visualisation of disparate heterogeneous data from multiple sources. Furthermore, the use of semantic and Linked Data technologies enables the detection rules to evolve along with fraudulent cases and thus to reflect the maturing of domain expertise and the evolution of intelligent society.

3. Fraud detection scenario

With respect to detecting hidden anomalies, we focussed on fraudulent claim cases in public councils as a scenario to pilot our technology. A scenario was developed by working closely with Fujitsu UK & Ireland and by soliciting requirements from customers.

Council social benefit handling departments have reported several categories of fraudulent claims including ID theft, housing benefit fraud, and council tax exemption, all of which could be tracked by monitoring and analysing various patterns in single and/or multiple data sets. Patterns of interest include, for example, "one mobile number registered for housing benefit claim in multiple councils" and "the distance between home address and school address of the claimant's

children is more than 50 miles." Councils often maintain a blacklist of existing fraudulent claims that can be checked and matched against new claims entering the system. Moreover, large volumes of data are continuously being added to the system, and such data needs to be subjected to semantically enriched anomaly detection in order to cope with this dynamism.

Benefit fraud detection is essentially a semantic alignment and pattern matching problem. Our studies have revealed several common heuristics that civil servants use to identify suspicious claims involving identity theft, housing benefit fraud, and council tax exemption fraud. Upon receiving a new claim, the typical process is 1) matching the claimant's name against those on the blacklist, 2) pooling data from multiple repositories (e.g., DVLA for vehicle registration, EduBase for school addresses and name disambiguation, and Land Registry for property addresses/coordinates), 3) aligning the claimant name with those on approved claims, 4) identifying potentially fraudulent claims, and 5) investigating them. It is evident that semantic technologies can play an important role in this process. In practice, ontologies and Linked Data are considered to be semantic methodologies while the RDF together with its query language SPARQL[13] and Web Ontology Language (OWL)[14] are typical examples of semantic technologies.

Semantic technologies enable the syntactic disguises used by fraud perpetrators to be stripped away. For instance, ontology engineering and the RDF can be used to identify relationships among claims. RDF represents data in triples: <*s,p,o*>, where *s* is the subject, *p* is the predicate, and *o* is the object. These triples can be used to capture the links between claimants on the one hand and their properties/assets, their family members, and their businesses/revenues on the other hand. The performance of the matching and reconciliation methods used to compare the attributes of new claimants with those of ones on the blacklist can be enhanced by using semantics. Inspecting new and existing claims involves semantic alignment and integration across multiple data sources. SPARQL queries and rules help to increase the coverage through semantic query rewriting and transformation. Detected anomalies can be clarified by using semantic equivalency relationships (e.g., <*owl:sameAs*> and <*skos:relatedMatch*>). Thus, by using semantic technologies, organisations can dynamically describe new fraud cases and facilitate the integration, analysis, and visualisation of disparate and heterogeneous data from multiple sources. The use of semantic technology to generate semantic fraud detection rules helps to convert labor intensive tasks into (semi-)automated processes.

4. Developed technology

For the social benefit fraud detection use case, we extracted the design requirements and designed a pilot system accordingly. The system architecture is illustrated in **Figure 1**. It combines practices common in the semantic technology community with characteristics specific to the fraud detection domain. All the data is duplicated in a graph data storage structure so that the integrity and ownership of the original data sources is maintained. The interface used to triplify data from existing RDBs is based on that of other systems targeting similar use cases. It is particularly useful for querying data in third-party RDBs and file-based systems. Data is triplified using a lightweight ontology manually crafted and based on existing models from different councils. Many open source tools are available for RDB-to-RDF format conversion (e.g., Google Refine with RDF extension).[15] New claims continuously arrive, and a consistent approach is used when updating the claim repository and the status of filed claims.

4.1 Data pre-processing

Data from a variety of sources and in a variety formats are incorporated into the fraud detection system with the help of an importing and preparation module responsible for format conversion, data alignment, data validation, triplification, annotation, etc.

The graph data structure underlying the BigGraph RDF data model facilitates the use of graph analysis algorithms to discover new fraud patterns and rules. Many such algorithms are either not readily available or are very expensive to implement in conventional tabular data models in RDBs. Graph data representation gives the system great flexibility and extensibility. In fact, the BigGraph platform is extensible in different directions. With the help of rapid, on-demand data reconciliation, data sets from different councils can be added or removed when deemed useful. At a fine granularity level, particular types of historical claims and some of their attributes can be selected or

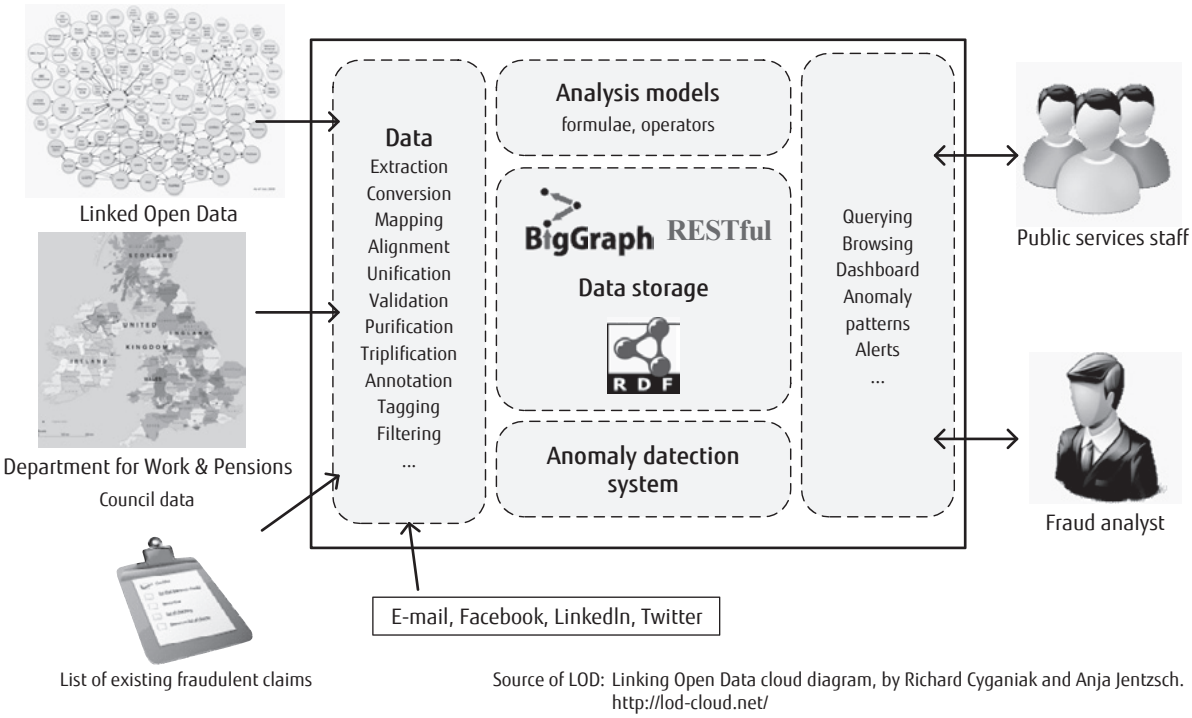

Figure 1
System architecture.

deselected to make certain characteristics stand out. It is also possible to change the matching query patterns of business rules to reflect newly acquired fraud knowledge. These changes can be made on the fly when users are experimenting with different scenarios, highlighting the benefit of the graph data structure.

Instead of incorporating many (potentially useful) data sources, we take an on-demand approach in order to maintain the balance between overhead and benefit: only the most necessary pubic data repositories are used to minimise the cost of data cleansing and conversion. Due to this consideration and customer requirements, many "fancier" data sets (e.g., London Gazette[16]) and news streams) are currently not used.

Data models from different councils present only simple semantic and syntactical discrepancies. Lightweight and supervised reconciliation methods based on string comparison algorithms are sufficient for model alignment. Reconciliation methods based on a graph structure are excluded due to their complexity, which could lead to significant delay when large amounts of claims arrive simultaneously.

4.2 Storing and accessing data

There are two types of claim data: the incoming flow of new claims and existing claims. The underlying storage system needs to accommodate both types efficiently.

The amount of triplified data for benefit fraud detection can potentially be very large, so a distributed ordered key value store (KVS) with enhanced range query and locality is used as the underlying RDF storage.[17] RDF triples are stored as keys of the key-value pairs while the value part is reserved for triple metadata (provenance trace, access control policy, caching, etc.). The storage can be deployed on the Global Fujitsu Cloud Platform.[18] Jena Graph is pipelined with OSS Hadoop HBase, resulting in scalability and transaction-oriented data operation through KVS and graph operations and RDF-specific reasoning through Jena. Jena represents the claim data as semantic graphs, enabling easy navigation through the data space by traversing along property and instantiation references.

The system interfaces with both human users and applications. The Web-based user interface is illustrated in **Figure 2**. A stream of new claims is dynamically visualised on the user interface with

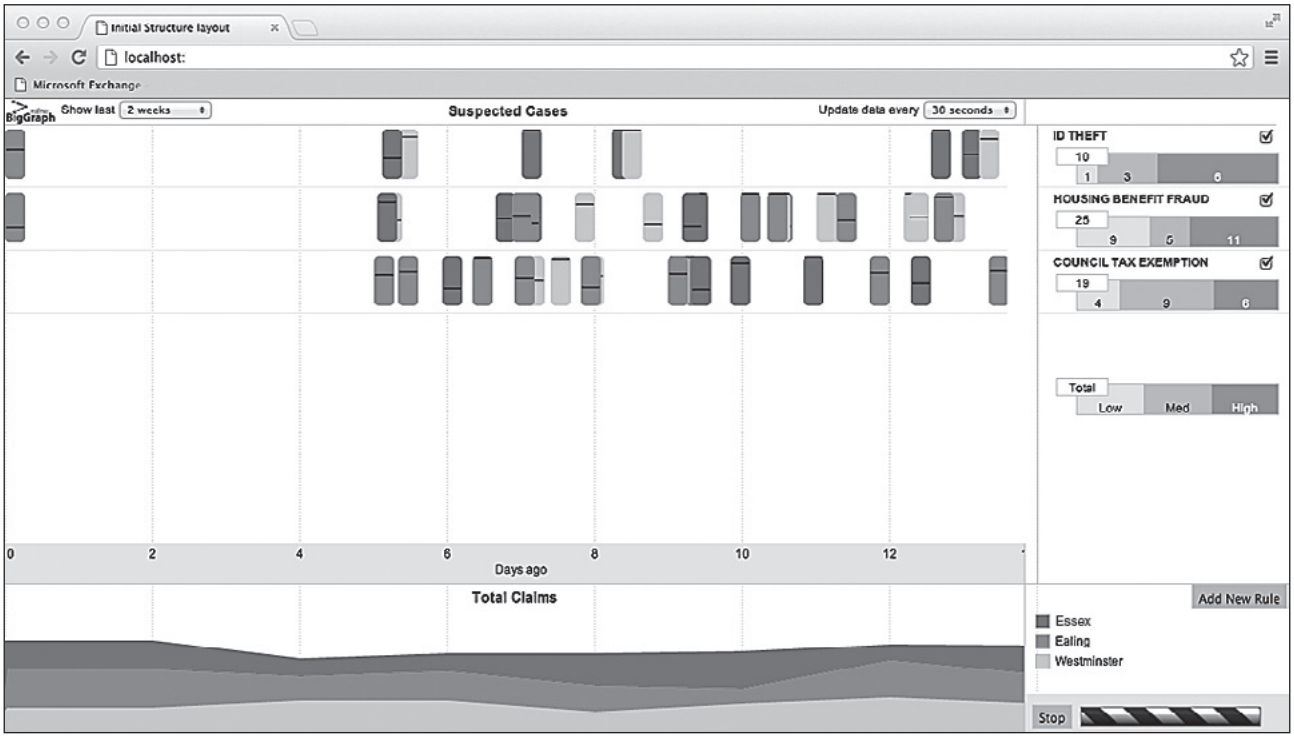

Figure 2
Web-based user interface.

color-coded indication of the data sources and confidence levels. A claims accumulation chart is displayed on the right side to give users an overview of the task backlog. The programming interface is underpinned by Fujitsu's implementation of the REST (REpresentational State Transfer)[19] software architecture, which standardises the modeling, exposure, and referencing of data items in the repository.

4.3 Enabling anomaly detection using BigGraph

Our data analytics technology aggregates data, relationships, and processing in a unified manner and presents the resulting new knowledge as a single common abstraction. This aggregated domain model enables fraud services organisations to take action in real time. More specifically, a data-centric and event-driven programming paradigm is used, and data analytics is attached and localised to individual data items (in terms of RDF data graph vertices). BigGraph uses a technology called "embedded local functionality" in which behaviours that are local processing units are attached to the data. They ensure actions and analytic outcomes are fully distributed to maximise parallelism.

BigGraph handles various types of information, processes external sources of data as a stream, and generates new knowledge, enabling hidden and fraudulent anomalies in Big Data to be detected. Anomaly detection is facilitated by event-based inference. Known fraud patterns are obtained from domain experts and coded as SPARQL rules. Rule patterns are matched against a semantic data graph, and multiple pattern expansion methods are used to maximise detection of possible anomalies. For instance, semantic alignment is performed to avoid ruling out potential matches; third-party ontologies (e.g., the GeoName ontology) are consulted for terminological knowledge.

5. Conclusion

As we have described, linking multiple data sets can reveal patterns and correlations that are otherwise locked in data silos. Furthermore, semantic technologies can be used to enhance social benefit fraud detection. For the fraud detection use case presented, the unique requirements from the domain were identified and an appropriate system was designed.

The BigGraph project focuses on how Fujitsu's anomaly detection technology can be applied to real-life problems so as to convert labor intensive tasks into (semi-)automated processes. BigGraph, as a data management platform, has distinctive advantages over existing solutions along all three Big Data axes.

1) BigGraph enables linking multiple data sources that are in different geographic locations and that use different models and different formats. Data *variety* is handled by links and relations in the underlying graph data representation.
2) BigGraph uses data/process distribution and data/process locality. Data *velocity* is therefore accommodated by distributed analytics acting locally on data items.
3) BigGraph handles data *volume* requirements by using distributed storage with full data replication and fault tolerance tuned in accordance with the RDF data model.

The presented use case demonstrates that BigGraph can be used to address real-life requirements and attack real-life problems. It provides a valuable alternative to RDBs for harnessing the true value of data in a fraud detection case where speed in analytics and insight obtained from data connections is increasingly critical. The current BigGraph implementation can be reinforced with richer data analytics and higher dynamism (e.g., on-the-fly pattern specification and rule definition).

From a broader perspective, BigGraph offers the means for public and private institutions to confront the Big Data challenge. It can help to extract patterns and trends from seemingly disordered, irrelevant data sets and to equip CIOs with a new apparatus for tackling long-standing data issues (e.g., availability, interoperability, sharing, reuse, and integration) inherent in the enterprise architecture without high capital expenditure, yet with good return on investment.

References

1) Z. Paul et al.: Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data (1st ed.). McGraw-Hill Osborne Media (2011).
2) A. Nasser et. al.: Enterprise BigGraph. 46th Hawaii International Conference on System Sciences. pp. 1005–1014, 2013. *http://origin-www.computer.org/csdl/proceedings/hicss/2013/4892/00/4892b005.pdf*
3) W3C: RDF (publication date: 2004-02-10). *http://www.w3.org/RDF/*
4) The Re-use of Public Sector Information Regulations (2005). *http://www.legislation.gov.uk/uksi/2005/1515/pdfs/uksi_20051515_en.pdf*
5) N. F. Authority: Annual fraud indicator (March 2012).
6) A. Mie: Government to crack down on welfare fraud as payouts balloon (June 2012). *http://world.topnewstoday.org/uk/article/2445084/*
7) T. Prenzler: Welfare fraud in Australia: Dimensions and issues. Australian Institute of Criminology (June 2011). *http://www.aic.gov.au/publications/current%20series/tandi/421-440/tandi421.html*
8) V. Chandola et al.: Anomaly detection: A survey. *ACM Comput. Surv.*, Vol. 41, Iss. 3, Article 15, (July 2009). DOI=10.1145/1541880.1541882. *http://doi.acm.org/10.1145/1541880.1541882*
9) L. Akoglu et al.: Anomaly, event, and fraud detection in large network datasets. The Sixth ACM International Conference on Web Search and Data Mining (WSDM '13). ACM, New York, NY, USA, pp. 773–774, 2013. DOI=10.1145/2433396.2433496. *http://doi.acm.org/10.1145/2433396.2433496*
10) B. Hu et al.: Applying Semantic Technologies to Public Sector: A Case Study in Fraud Detection. Semantic Technology. Lecture Notes in Computer Science, Vol. 7774, pp. 319–325, (2013).
11) DataGov: The principles of open public data (June 2010). *http://data.gov.uk/library/public-data-principles*
12) E. Thomsen: OLAP Solutions: Building Multidimensional Information Systems. John Wiley & Sons (1997).
13) W3C: SPARQL Query Language for RDF. *http://www.w3.org/TR/rdf-sparql-query/*
14) W3C: OWL Web Ontology Language Overview. *http://www.w3.org/TR/owl-features/*
15) DERI: RDF Refine. *http://refine.deri.ie/*
16) London Gazette: Gazettes. *http://www.london-gazette.co.uk/*
17) B. Hu et al.: Towards big linked data: a large-scale, distributed semantic data storage. The 14th International Conference on Information Integration and Web-based Applications & Services (IIWAS '12). ACM, New York, NY, USA, pp. 167–176, 2012.
18) Fujitsu Global Cloud Platform. *http://www.fujitsu.com/fts/cloud/solutions/global-cloud-platform/*
19) L. Richardson and S. Ruby: RESTful Web Services. O'Reilly Media, 2007.

Bo Hu
Fujitsu Laboratories of Europe Limited
Mr. Hu is engaged in research and development of Linked Data solutions for healthcare, business intelligence, and intelligent society.

Aisha Naseer
Fujitsu Laboratories of Europe Limited
Ms. Naseer is engaged in research and development of Big Data and a Linked Data platform.

Takahide Matsutsuka
Fujitsu Laboratories of Europe Limited
Mr. Matsutsuka is engaged in development of large-scale distributed databases and processing platforms utilizing Linked Data.