Algorithmic tools for data-oriented law enforcement

Cocx, T.K.

Appendix A

Statistical Significance and Privacy Issues

An intrinsic issue with the approaches described in this thesis is that there is a significant privacy and reliability issue, especially when they are applied on data as sensitive as that collected in the process of law enforcement. Applicability in daily police practice depends on the statistical insightfulness and relevance, current law and privacy regulation. In this chapter we describe some of the features of our approaches in this context and elaborate on their value in the application domain of law enforcement.

A.1 Statistics

A first important notion in the discussion of statistical issues surrounding our data mining approaches is that most of our efforts involved an *unguided search*, meaning that there was no specific occurrence or relation that we directed our searches toward. An advantage to this way of examining raw data is that more and especially potentially unexpected results are retrieved more easily than in a directed, in-depth statistical analysis. An intrinsic drawback to these kind of approaches, however, is that its results are founded on some statistical occurrence with a certain reliability, but are not able to, nor strive to, provide an explanation concerning the existence of the discovered patterns and in particular not on any underlying causality.

Due to this nature of the methods that are the algorithmic core of our approach, our tools and the accompanying results are less suitable in the research area of the social sciences, where insights into the background, existence and causality in the emergence of certain patterns are the main objective of scientific efforts. However, algorithms like the ones described in this thesis have been applied within a large number of different application domains, mostly in the form of decision support systems. Although one of the most usual deployment of these techniques is in the retail industry, both for companies and their customers, they are applicable in almost any area where decisions are to be

made and where large amounts of data are available to have computer guided statistical data assistance. As is shown in the Introduction, some of these applications are currently used in the area of law enforcement, even though the principles underlying the emerging patterns are not revealed by the tools themselves. In cases where such an explanation is needed, the data resulting from the unguided search should, and is usually examined by a domain expert.

A good example where computer guided statistical data analysis is successfully implemented in the domain of law enforcement is the detection of fraudulent monetary transactions. Unguided searches through large database of transactions have been used widely in the process of discovering money laundering activities or tax evasion. If a pattern emerges, it is examined by a domain expert and appropriate action is taken according to his or her findings.

Below we provide a chapter wise discussion of the statistical applicability of the respective tools, that are an important background when the systems are to be put into actual use.

In Chapter 2 we describe a tool that yields relations (patterns) between the co-occurrence of crimes and demographic information. In this case, the pattern frequency is denoted by the percentual co-occurrence of its member in the data set. The novelty in this chapter resides in the way that these patterns are retrieved rather than a difficult statistical construct. The emerging patterns, accompanied by their frequency, can therefore be taken "as they are", describing a purely factual statement of occurrence. The retrieval methods do not explain why these patterns emerge, but the patterns themselves can be used to shed light on other existing phenomena within crime. The statistical means to acquire the patterns are widely known and applied. The same situation is applicable in Chapter 6, where the main focus is on the fast construction of the frequent subcareers, but the quality of the results is measured by a simple count of occurrences.

The visualization methods discussed in Chapter 3 are a good example where immediate feedback is provided about the statistical error that is made during the image construction process. The user is constantly informed through color coding how reliable the constructed visualization is, by inspecting the global error. This method is also used as an integral part of the approach in Chapter 4, where the visualization is the main result of the tool, accompanied by a clustering underlying it.

The distance measures discussed in Chapter 5 are all widely applied and their respective reliabilities and likelihood ratios are well-known. They are already used in many areas where reliability and robustness is of the essence, e.g., medical sciences, forensic research, however, the visualization method is quite different from the one used in Chapter 4. Therefore, a graph displaying the global error is displayed within the tool to inform the user of the quality of the constructed visualization (and therefore its underlying clustering). This is also the case for the construction of the visualization in Chapter 9. In that chapter the calculation of the distances is based upon the regular normal distribution.

In Chapter 7 and Chapter 8 a way of predicting the continuation of criminal careers is discussed. Since the possible outcome of such tools is highly sensitive we calculate two reliabilities. The first reliability describes how many times the prediction yields an accurate response in all cases known up to the current moment, providing users with a

"general" reliability of the method. The second calculation is used to provide a reliability to a single prediction currently under consideration. The error made in the reliability calculation is 0.1, which is seen as very reliable. The same accuracy is reached in Chapter 10 if the most strict rules are used for set-separation.

In general, our methods are all subject to errors, but this error is assessed and made available to the user in every case. Especially since our tools do not describe the underlying principles for the patterns they yield, we encourage every possible user to be familiar with the statistical material described in the respective chapters of this thesis and familiarize themselves with the figures describing the reliability of the outcome.

A.2 Privacy

Naturally, approaches such as in this thesis come with privacy concerns. Authorative works on the application of data mining in the area of law enforcement in the Netherlands (our application domain, both through residence and project cooperation) are [37, 36], that discuss most relevant issues in law.

Every time data is analyzed for the purpose of law enforcement, the personal privacy is violated in a formal sense. According to international and national law, this is only acceptable if sufficient judicial principles exist that are known and accessible by civilians. It is important that laws are exact, so that civilians know beforehand when and how law enforcement agencies are allowed to use their authority in this sense. From this point of view, there are two main principles:

- Data on unsuspected individuals should only be analyzed in situations that are highly exceptional.
- Data on individuals should only be analyzed for the purpose it was collected for.

In the Netherlands these matters are arranged in *WPolr*(Law Police Registers) as a special case of the *Wbp* (Law Protection Personal Data). Important aspects in the application of these laws are the distribution of data and the principle of necessity. According to the first, the only police registers that can be distributed and used for most police tasks are the general police registers. According to the latter, every gathering, storage and analysis of data should be necessary for the police task. Each police register should have attached rules about how these principles are met [37]. The database and its current usage [29], described in Appendix B, comply with these demands.

Most of our methods are in complete accordance with these laws since they are drawn from existing police registers, meant for statistical analysis of trends in crime and should and can only applied in the area of crime analysis.

However, the tools described in Chapter 7 and Chapter 8 deal with the translation of newly discovered patterns or descriptions to individual cases rather than a general quest for truth or the discovery of general knowledge. This approach is best compared to the process of discovering suspicious financial transactions like for example in [23]. These methods also enforce criteria from a knowledge discovery support system on individual financial transactions in order to find outliers that might indicate for example money

laundry activities. Naturally, not all transactions that yield warnings are actually fraudulent, but as long as the expected chance a warning is actually correct is reasonably high, the usage of such an decision support system is warranted as long as each warning is reviewed by experts before action is undertaken. The same should apply to the usage of our approach. However, in contrast with, among others, the above mentioned financial monitoring systems, that monitor *all* transactions done by *everybody*, only the people who have been active criminals for more than three years are under surveillance and even then, they are only electronically processed every time they commit a new crime.

However, our approach and the search for fraudulent transactions differ in one thing. The necessity principle has already been established for the latter, while no such directive has been determined for our tools. It is questionable if the necessity for suspect monitoring on possible criminal careers currently exists, especially since it is unknown if crime can be decreased though usage of our tools. Therefore, our tools in this matter are probably currently not allowed in the area of law enforcement, nor are they expected to be in the near future.

In Chapter 10 we describe a tool that uses publicly available data, to determine if there are aspects to users of Social Networking Sites that make them fall into danger categories for online predatation. It may however serve its purpose as a monitoring tool for larger internet applications, identifying profiles that may pose a risk to children and could be reviewed by a detective, but the possibilities in this area heavily depend on judicial constraints within the country of origin or application.

In the Netherlands, such a usage of this tool probably violates the first principle of the WPolr; our tool gathers data on unsuspected individuals, without the existence of a highly exceptional situation. Even though the data is publicly available and the data is not necessarily stored in a (semi-)permanent police register, the data is processed and therefore the usage of the tool in such a way is probably currently not possible or warranted.

However, in contrast to the more active approaches described in that chapter, the method has also potential as a strategic tool, helping to chart the dangers of predators on Social Networking Sites that could influence future legislation or police priorities.

Within this thesis we described methods that show the potential for computer guided statistical analysis on police data. However, usage and applicability are regulated by law, daily police activities and social necessity. Therefore, potential users are advised to familiarize themselves with the law in these matters before proceeding with the research in this thesis. Since in a chain of command, the situation might exist that executive personnel is not familiar with legislation concerning their activities and management personnel is not aware of the exact nature of the tasks their subordinates perform, the tools discussed in this thesis should provide proper warning of the privacy sensitive nature to any potential user, that might not be familiar with legislative concerns on their usage.

During our research no “new” potential offenders were identified nor put in any list that is currently monitored by police officials. Also, all data used within the research was either publicly available on the internet or made available by the police in an anonymized version. Hence, we assume no privacy sensitive footprints of our research remain.

Appendix B

The HKS database

A number of chapters deal with methods that are either designed to work with or tested on a database of criminal activity, describing offenders and their respective crimes. Throughout our research, we used the National HKS database of the Dutch National Police as a representative example. In this appendix we describe the structure of this database and the terms of use we observed during the testing phases of our research.

B.1 Structure

The Dutch National Police (Korps Landelijke Politie Diensten, KLPD), through its National Detective Force Information Department (Dienst Nationale Recherche Informatie, DNRI), annually extracts information from digital narrative reports stored throughout the individual, regional, administrative departments, and compiles this data into a large and reasonably clean database that contains all suspects and their possible crimes from the last decade. Since it is drawn from data stored in the individual HKS (Recognition systems, HerKenningsdienstSystemen), it is referred to as the National HKS Database.

Through a join on the municipal administration number (Gemeentelijke Basis Administratie, GBA) of the suspect, demographic information from Statistics Netherlands (Centraal Bureau voor de Statistiek, CBS) was added to the National HKS, for example including the age a person first committed a crime, his/her nationality and (ethnic) descend, and the perpetrator's gender [7, 36, 29].

In the main tables of the National HKS there are approximately one million rows, representing suspects and approximately 50 columns representing offenses and demographic data. All crimes are stored in a separate table that contains more details on this event (i.e., crime date, type) and can be linked to the individuals in the larger table. Crimes of an individual are split up into eight different types, varying from traffic and financial infringements to violent and sex crimes. All these crimes have got an intrinsic seriousness attached to them.

The compilation of the database started in 1998 and therefore only contains suspects and their potential offenses that have been reported within of after this year. However,

every individual that is reported in this period is stored in the database with all its "baggage", meaning that all potential offenses for this individual are present in the National HKS, even if they were reported before 1998.

The National HKS is stored in two distinct forms: the unmodified database as compiled from the individual departments and CBS, and an anonymized version where all identifiable information on individuals is omitted and replaced by a unique "suspect-number".

B.2 Ownership and Residence

The ownership of the National HKS is rather complex and subtle, for the most part because there are many original sources, e.g., all Dutch municipalities and all local Dutch police departments, but for all practical reasons administrative ownership lies with both CBS and KLPD, where the latter is responsible for and owner of the characteristic part, e.g., the data on crimes. The ownership of data is depicted in Figure B.1

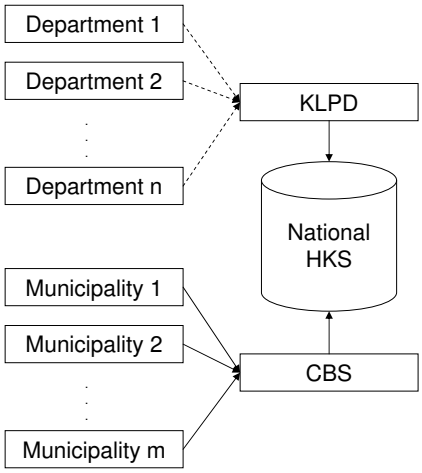

Figure B.1: Ownership of the National HKS

The actual situation, however, is somewhat more complex. Although original ownership on the narrative reports stored in the HKS lies with the individual police departments, they are not allowed to store data on suspects after a certain expiration date that is based upon closure of the case or some preset time frame set by the law (cf. Appendix A). The KLPD, however, is authorized to backup the data for a longer time if they restrict its usage to the general study of crime, its development in the country and offenders. By principle, the regional departments (or the KLPD) is not allowed to use any data on specific individuals that reside in the National HKS but is no longer available in any separate, regional HKS. Therefore, the ownership of the compiled data from regional departments can no longer be said to reside with these departments when the original data has been

deleted from their systems, but is transferred to the KLPD, denoted by the dotted lines in Figure B.1.

There are two main locations where the National HKS is stored and accessible for usage: the headquarters of CBS and the Zoetermeer location of the KLPD, under auspices of its DNRI department. The copy at CBS is, as a principle, accessible by all who want to study it for scientific purposes, but only resides at their location in the anonymized version. The KLPD copy, on the other hand, is only accessible by KLPD personnel or by special request that is well motivated, especially since they manage both the original and anonymized version. The demographic data portion of the National HKS is available without any restriction in both locations.

B.3 Common Usage

There are two main objectives for the compilation of the national HKS: the possibility to answer general questions on the nature of or trends within crime to law enforcement officials and the compilation of the Nation-wide Crime Chart (Landelijke Criminaliteits-Kaart, LCK [29]). The first objective is met by the ability to contact the DNRI by phone.

The LCK is an annual report, fabricated by DRNI, that describes a wide variety of data and knowledge on and trends within crime and observes a large number of trends within the crime area, providing ample insights into the relations between crime, city-size, ethniticity, law enforcement policy and police deployment.

Although they are consistently referred to as suspects, all individuals are treated as offenders in these publications, as can be derived from nomenclatures like “recidivism”, “criminal foreigners” and “criminal” profiles, and all narrative reports are considered to describe events that really happened. No checking is performed on the outcome of police investigations or court proceedings, nor are they considered relevant in the pursuit of the objective to provide insights into the existence and development of crime.

Apart from the two owning organizations, research on the National HKS is also done by the regional police departments, ministries, the Netherlands Court of Audit (Algemene Rekenkamer), municipalities and research organizations, mostly in the anonymized form.

B.4 Usage during Research

During the research underlying this thesis, we performed a number of tests on a collection of meta-data derived from the National HKS. For this purpose we obtained permission from the KLPD to use their copy of the anonymized version, provided the database remained at KLPD headquarters.

Before the data was analyzed by our tool it was transformed into numbered series of integers that represent criminal careers, but were untraceable to the original data. Two steps were taken within this encoding:

- **No numbering on suspects** Since our research does not deal with any individual suspects, nor do we need information uniquely identifying them, all unique numbering was removed.

- **Offset all time frames** Since we are not interested in data on actual time frames, relating certain events to certain real life periods, we offset all data on this to start at time frame 1, identifying the first time frame of activity that is increased with 1 for each next time frame.

An example of this can be viewed in Figure B.2, where the numbers represent specific crime types.

Year Suspect	2002	2003	2004	2005
23		15 15	38	15 38
25	15		29	

Transformed to

1:	1:	15
		15
	2:	38
	3:	15
		38
2:	1:	15
	3:	29

Figure B.2: Fictive example of the transformation of HKS data before analysis by our tools

As is clear from the figure, all relation between original suspects and numbered series are lost. On top of that, there is no longer a relation between actual years and the numbered time frames in the transformed list, especially since there is no relation between the first year of activity for suspect 1 and the first year of suspect 2 (both represent different years but are denoted by 1).

Just as in the official KLPD publications (like the LCK), we assumed all suspects were actual offenders and all reported events had actually happened. However, no conclusions were drawn about specific or retraceable individuals. Also, all data resulting from our research on the National HKS has been made available to the KLPD personnel attached to our project for further examination.