Criminal and Civil Identification with DNA Databases Using Bayesian Networks

Marina Andrade marina.andrade@iscte.pt
Department of Quantitative Methods
ISCTE – Lisbon University Institute
Lisbon, 1649-026, Portugal

Manuel Alberto M. Ferreira manuel.ferreira@iscte.pt
Department of Quantitative Methods
ISCTE – Lisbon University Institute
Lisbon, 1649-026, Portugal

Abstract

Forensic identification problems are examples in which the study of DNA profiles is a common approach. Here we present some problems and develop their treatment putting the focus in the use of Object-Oriented Bayesian Networks - OOBN. The use of DNA databases, which began in 1995 in England, has created new challenges about its use. In Portugal the legislation for the construction of a genetic database was defined in 2008. With this it is important to determine how to use it in an appropriate way.
For a crime that has been committed, forensic laboratories identify genetic characteristics in order to connect one or more individuals to it. Apart the laboratories results it is a matter of great importance to quantify the information obtained, i.e., to know how to evaluate and interpret the results obtained providing support to the judicial system. Other forensic identification problems are body identification; whether the identification of a body (or more than one) found, together with the information of missing persons belonging to one or more known families, for which there may be information of family members who claimed the disappearance. In this work we intend to discuss how to use the database; the hypotheses of interest and the database use to determine the likelihood ratios, i.e., how to evaluate the evidence for different situations.

Keywords: Bayesian networks, DNA profiles, identification problems.

1. INTRODUCTION

The use of networks transporting probabilities began with the geneticist Sewall Wright in the beginning of the 20th century (1921). Since then their use had different forms in several areas like social sciences and economy – in which the used models are, in general, linear named Path Diagrams or Structural Equations Models (SEM), and in artificial intelligence – usually non-linear models named Bayesian networks also called Probabilistic Expert Systems (PES), [11],[14].

Bayesian networks are graphical structures for representing the probabilistic relationships among a large number of variables and for doing probabilistic inference with those variables, [13]. Before we approach the use of Bayesian networks to our interest problems we briefly discuss some aspects of PES in connection with uncertainty problems in section 2.

In section 3 after presenting some possible forensic identification problems the creation and use of DNA profile databases in some European countries is discussed putting the focus in the entry criteria and the differences observed. How to approach evaluate and interpret the results, whether it is a criminal or a civil identification problem is presented in section 4. Thus, it is important to describe the Portuguese law establishing the principles to maintain a DNA database file for civil and criminal identification purposes. The study of a criminal identification problem considering one single perpetrator, and a civil identification problem with one volunteer, for two different situations, are considered.

2. EXPERT SYSTEMS

Expert systems are attempts to crystallize and codify the knowledge and skills of one or more experts into a tool that can be used by non-specialists, [14]. An expert system can be decomposed as follows:

Expert system = knowledge base + Inference engine.

The first term on the right-hand side of the equation, knowledge base, refers the specific knowledge domain of the problem. The inference engine is given by a set of algorithms, which process the codification of the knowledge base jointly with any specific information known for the application in study.

Usually it is presented in a software program, as the one we are going to show hereafter, but such is not an imperative rule. Each of those parts is important for the inferences, but knowledge base is crucial. The inferences obtained depend naturally on the quality of the knowledge base, of course in association with a sophisticated inference engine. The better those parts are the best results we can get.

A PES is a representation of a complex probability structure by means of a directed acyclic graph, having a node for each variable, and directed links describing probabilistic causal relationships between variables, [1]. Bayesian approach is the adequate for making inferences in probabilistic expert systems.

2.1 Bayesian networks

Bayesian networks are graphical representations expressing qualitative relationships of dependence and independence between variables. A Bayesian network is a directed acyclic graph $\mathcal{G}$ (DAG) having a set of V vertices or nodes and directed arrows. Each node $\upsilon \in V$ represents a random variable X_υ with a set of possible values or states. The arrows connecting the nodes describe conditional probability dependencies between the variables.

The set of parents, $pa(\upsilon)$, of a node υ comprises all the nodes in the graph with arrows ending in υ. The probability structure is completed by specifying the conditional probability distributions for each random variable X_υ and each possible configuration of variables associated with its parent nodes $x_{pa(\upsilon)}$. The conditional distribution of X_υ is expressed given $X_{pa(\upsilon)} = x_{pa(\upsilon)}$. The joint distribution is $p(x) = \Pi_{\upsilon \in V}\, p(x_\upsilon | x_{pa(\upsilon)})$. There are algorithms to transform the network into a new graphical representation, named junction tree of cliques, so that the conditional probability $p(x_\upsilon | x_A)$ can be efficiently computed, for all $\upsilon \in V$, any set of nodes $A \subseteq V$, and any configuration x_A of the nodes X_A. The nodes in the conditioning set A are generally nodes of observation and input of evidence $X_A = x_A$, or they may specify hypotheses being assumed.

Software such as Hugin[1] can be used to build the Bayesian network through the graph $\mathcal{G}$. That can be done by specifying the graph nodes, their space of states and the conditional probabilities $p(x_v| x_{pa(v)})$. In the compiling process the software will construct its internal junction tree representation. Then, by entering the evidence $X_A= x_A$ at the nodes in A, and requesting its propagation to the remaining nodes in the network, the conditional probabilities $p(x_v| x_A)$ are obtained. Version 6.4 of Hugin and upgrades allow the graphical use of OOBN.

OOBN are one example of the general class of Bayesian networks. An instance or object is a regular network possessing input and output nodes as well as ordinary internal nodes. The interface nodes have grey fringes, with the input nodes exhibiting a dotted line and the output nodes a solid line. The instances of a given class have identical conditional probability tables for non-input nodes. The objects are connected by directed links from output to input nodes. The links represent identification of nodes. We use bold face to refer the object classes and math mode to refer the nodes. The modular flexibility structure of the OOBN is of great advantage in complex cases

3. Forensic identification problems

The use of DNA profiles in forensic identification problems has become, in the last years, an almost regular procedure in many and different situations. Among those are: 1) disputed paternity problems, in which it is necessary to determine if the putative father of a child is or is not the true father; 2) criminal cases as if a certain individual Y was the origin of a stain found in the scene of a crime; or in more complex cases to determine if an individual or more did contribute to a mixture trace found; 3) civil identification problems, i.e., the case of a body identification, together with the information of a missing person belonging to a known family, or the identification of more than one body resultant of a disaster or an attempt. And even immigration cases in which it is important to establish family relations.

Here the focus is to approach the civil and criminal identification problems. The establishment and use of DNA database files for a great number of European countries worked as a motivation to study in more detail the mentioned problems and the use of these database files identification.

The use of a DNA profile database may allow delinquents' identification and/or the connection of criminal conducts and the respective individuals, the exclusion of innocents as well as the recognition and civil identification of missing people. A genetic profile database may be an important help in forensic investigation, particularly in crimes of repetitive tendency, when DNA samples of condemned individuals are collected. In the context of the civil identification it may be very useful when unidentified corpses appear and may be identified by comparison of their DNA profiles family volunteer's profiles.

3.1 DNA database files

The discovery of biological fingerprints in 1984 opened new perspectives in forensic identification area. Since then the technical advances and results obtained have allowed studying and solving increasingly complex problems. Almost twenty five years after Alec Jeffreys' team discovery, Portugal established the legislation for the construction of a genetic database, law nº5/2008.

The advances in DNA technology and knowledge opened new perspectives for civil and criminal investigation. Apart from the ethical and legal questions that are in the domain of the legal system, we should draw some attention to the experience and knowledge acquired by those countries that already have their own databases operating and try to learn from them ways to improve on how to operate with the Portuguese database, [5].

[1] http://www.hugin.com - OOBN a resource available in the Hugin 6.4 software.

Country	Year	Entry criteria for suspects (convicted offenders)
England	1995	Any recordable offence
Austria	1997	Any recordable offence
Croatia/Switzerland	2000	Any recordable offence
Germany	1998	> 1 year in prison (after court decision)
Finland	1999	> 1 year in prison
Denmark	2000	> 1.5 years in prison
Norway	1999	Many serious offences (after court decision)
Hungary	2003	5 years in prison
Sweden	2000	No suspects entered (> 2 years convicted)
Belgium	2002	No suspects entered (after court decision)
Netherlands	1997	No suspects entered (> 4 years convicted)
France	2001	No suspects entered (serious offences, voluntary samples only)
Spain	1998	Phoenix program – civil database for civil ident. vol. donations
Portugal	2008	Vol.,“problem samples”, “reference samples” (≥ 3 years convicted)
Italy	-	Law in preparation

TABLE 1: National DNA databases in Europe.

There has been a considerable discussion about the individuals to include in a DNA profiles database, specially with different results in countries with different legal systems. The main differences are linked to the emphasis given by the countries: to the individuals or the social order.

In the table above (TABLE 1) it is possible to observe significant differences between the European countries in what concerns criteria to enter a person into a database (while a suspect or only after conviction, different types of conviction). The criteria to remove records and the number of entries in the database also have important differences all over Europe.

As we have seen there are clear differences between countries more on the north and more on the south of Europe, which is essentially due the different perspective valuing more the social order or the individual itself.

4. Criminal and civil identification using DNA profile databases

Let us consider a criminal case in which a DNA profile has been recovered from a crime scene, and it is admitted to be left by the culprit (only one perpetrator); and a civil identification problem with one volunteer giving his/her genetic information.

The Portuguese law n°5/2008 establishes the principles for creating and maintaining a database of DNA profiles for identification purposes, and regulates the collection, processing and conservation of samples of human cells, their analysis and collection of DNA profiles, the methodology for comparison of DNA profiles taken from the samples, and the processing and storage of information in a computer file.

Here it is assumed that the database is composed of a file containing information of samples from convicted offenders with 3 years of imprisonment or more - α; a file containing the information of samples of volunteers - β; a file containing information on the "problem samples" or "reference samples" from corpses, or parts of corpses, or thing or place where the authorities collect samples - γ.

4.1 Criminal identification – one single perpetrator

For a crime that has been committed, forensic laboratories identify genetic characteristics in order to connect one or more individuals to the crime. Apart from the laboratories results it is a matter of great importance to quantify the information obtained, i.e., to know how to evaluate and interpret the results obtained providing a support to the judicial decision. Here it is assumed a DNA trace found at a scene of a crime left by only one perpetrator.

“The experience with some databases seams to indicate that before the commitment of a serious crime, some suspects have already been involved in minor offences. This fact associated to the repetitive motif of some serious crimes can support the importance of DNA databases not only to the criminal investigation but also to the prevention of crime, mainly if there are large entry criteria.”, [4].

Some notation:
Let C_c be the genetic characteristic (DNA profile) found at the crime scene, and C_s the suspect’s genetic characteristic. The evidence is $E=(C_c, C_s)$, the DNA typing results for the crime sample and the suspect. Our interest is to discuss how to evaluate the hypotheses and the *odds* ratio. In court the hypotheses are: *H_P*: The suspect *(s)* left the crime stain *vs H_D*: Some other person left the crime stain.

There is a match between the crime scene profile and the suspect’s profile. The court wants to compare the two preceding hypotheses. It is important to discuss the presentation of the evidence in court and how to evaluate the hypotheses of interest. The *posterior odds* are:

$$\frac{P(H_P \mid E, s\in\alpha)}{P(H_D \mid E, s\in\alpha)} = \frac{P(E \mid H_P, s\in\alpha)}{P(E \mid H_D, s\in\alpha)} \frac{P(H_P \mid s\in\alpha)}{P(H_D \mid s\in\alpha)}$$

$$= \underbrace{\frac{P(C_c, C_s \mid H_P, s\in\alpha)}{P(C_c, C_s \mid H_D, s\in\alpha)}}_{LR} \frac{P(H_P \mid s\in\alpha)}{P(H_D \mid s\in\alpha)}$$

The likelihood ratio, LR, takes the form:

$$LR = \frac{P(C_c \mid C_s, H_P, s\in\alpha)}{P(C_c \mid C_s, H_D, s\in\alpha)} \frac{P(C_s \mid H_P, s\in\alpha)}{P(C_s \mid H_D, s\in\alpha)}$$

Whether or not the suspect left the crime sample that does not provide any information to our uncertainty about his/her genetic characteristic or genotype, i.e., $P(C_s|H_P, s \in \alpha) = P(C_s|H_D, s \in \alpha)$. Therefore

$$LR = \frac{P(C_c \mid C_s, H_P, s \in \alpha)}{P(C_c \mid C_s, H_D, s \in \alpha)}.$$

In a case with a trace of a single perpetrator, for each marker, one can use the object-oriented Bayesian (OOBN) network shown in Figure 1. Each node (instance) in the network represents itself a Bayesian network. Nodes **spg** and **smg** are all of class founder, a network with only one node which states are the alleles in the problem and the respective frequencies in the population, and represent the suspect's paternal and maternal inheritance. Node **sgt** and **cgt** are of class genotype. The genotype of an individual is an unordered pair of alleles inherited from paternal (*pg*) and maternal (*mg*) genes, here represented by *gtmin:=min*{*pg*, *mg*} and *gtmax:=max*{*pg*, *mg*}, where *pg* and *mg* are input nodes identical to the *gene* node of *founder*. The nodes **cmg** and **cpg** specify whether the correspondent allele is or is not from the suspect. If **s=c?** has true for value then the true perpetrator's allele will be identical with the suspect's allele, otherwise the true perpetrator's allele is chosen randomly from another man in the population. The single node **s=c?** represent the binary query 'Is the suspect the perpetrator?'

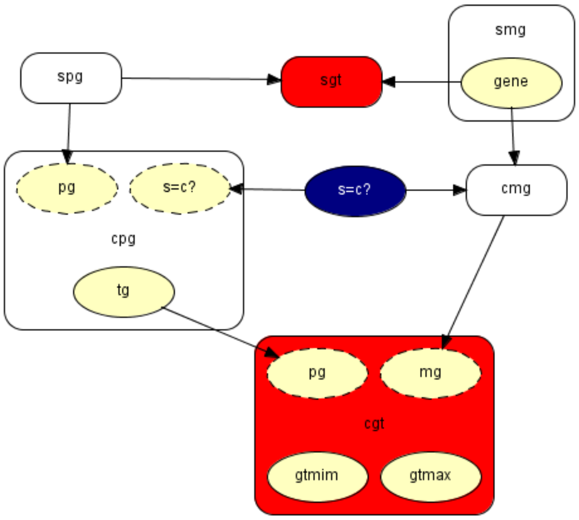

FIGURE 1: Network for a criminal case with a single perpetrator.

Here, if the suspect, s, is in the file of convicted offenders then $P(H_P \mid s \in \alpha) > P(H_D \mid s \in \alpha)$, otherwise those probabilities may be assumed equal. Therefore, the posterior odds are:

If s is in the file of convicted offenders Otherwise

$$\frac{1}{P(C_c \mid C_s, H_D, s \in \alpha)} \underbrace{\frac{P(H_P \mid s \in \alpha)}{P(H_D \mid s \in \alpha)}}_{>1} > \frac{1}{P(C_c \mid C_s, H_D, s \in \alpha)}.$$

We need to assess $P(H_P \mid s \in \alpha)$ and $P(H_D \mid s \in \alpha)$. The posterior odds computation is in the judges' domain, we can only explain how to do it and how to interpret the evidence, i.e., how to use it as a decision support element.

4.2 Civil identification – one missing person and one volunteer

A missing individual is reported. A body is found. The hypotheses are: *H*H_P: The body found is the body of the claimed individual *X vs* H_D: The body found is any other individual's, not the claimed individual *X*. People want their relatives alive and reject the hypothesis that states their lost relative is dead. A volunteer supplies genetic material to be used in the test of a partial match. The evidence is $E = (C_{Bf}, C_{vol})$ - the genetic characteristics of the body found, C_{BF}, and the volunteer, C_{vol}. The *posterior odds* is

$$\frac{P(H_P \mid E, vol \in \beta,\gamma)}{P(H_D \mid E, vol \in \beta,\gamma)} = \frac{P(E \mid H_P, vol \in \beta,\gamma)}{P(E \mid H_D, vol \in \beta,\gamma)} \frac{P(H_P \mid vol \in \beta,\gamma)}{P(H_D \mid vol \in \beta,\gamma)}$$

One may assume $P(H_P \mid vol \in \beta,\gamma) = P(H_D \mid vol \in \beta,\gamma)$ then

$$\frac{P(H_P \mid E, vol \in \beta,\gamma)}{P(H_D \mid E, vol \in \beta,\gamma)} = \underbrace{\frac{P(C_{BF}, C_{vol} \mid H_P, vol \in \beta,\gamma)}{P(C_{BF}, C_{vol} \mid H_D, vol \in \beta,\gamma)}}_{LR}$$

$$\frac{P(H_P \mid E, vol \in \beta,\gamma)}{P(H_D \mid E, vol \in \beta,\gamma)} = \frac{P(C_{BF} \mid C_{vol}, H_P, vol \in \beta,\gamma)}{P(C_{BF}, \mid C_{vol}, H_D, vol \in \beta,\gamma)} \underbrace{\frac{P(C_{vol} \mid H_P, vol \in \beta,\gamma)}{P(C_{vol} \mid H_D, vol \in \beta,\gamma)}}_{=1}.$$

The conditioning does not include the information of the body found. Whether or not the volunteer is related with the individual whose body was found that does not provide information to our uncertainty about his genotype.

Depending on the volunteer and the claimed individual family relation we only may observe a partial match with one volunteer. Apart form that, it is important to check if there is a match between C_{BF} and any of the "problem samples" in γ. Assuming no match of C_{BF} and any sample in γ, the likelihood ratio may be written as:

$$\frac{P(C_{BF} \mid C_{vol}, H_P)}{P(C_{BF} \mid H_D)}.$$

In a case having only one volunteer the likelihood ratio can be computed using a Bayesian network. Suppose we have a missing individual, an elder person and a son or a daughter claiming the disappearance and who gives a genetic profile voluntarily. The likelihood ratio can be computed using the following network:

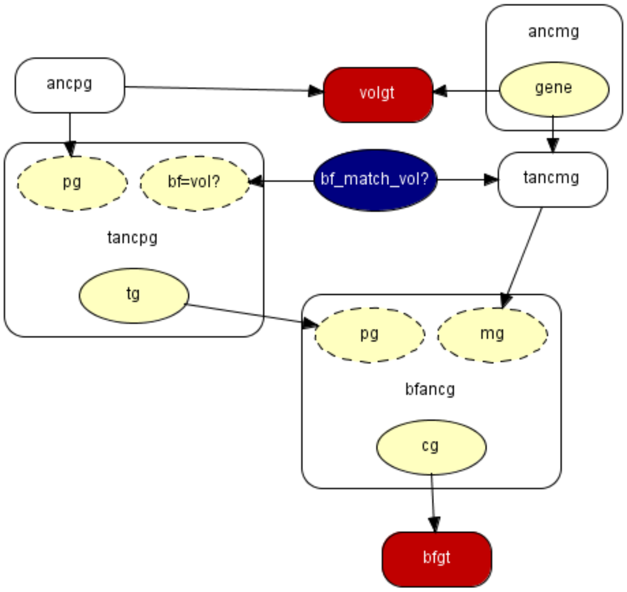

FIGURE 2: Network for civil identification with one volunteer, paternal or maternal relationship case.

As above nodes **ancpg** and **ancmg** are all of class founder, a network with only one node which states are the alleles in the problem and the respective frequencies in the population, and represent the volunteer's ancient paternal and maternal inheritance. Node **volgt** and **bfgt** are of class genotype, the volunteer and body found genotypes.

The nodes **tancmg** and **tancpg** specify whether the correspondent allele is or is not from the volunter. If **bf_match_vol?** has true for value then the volunteer's allele will be identical with the volunteer's allele. The node **bfancg** defines the Mendel inheritance in which the allele of the individual whose body was found is chosen at random from the ancient's paternal and maternal gene.

For a case with only one volunteer, but now for example a brother or a sister of an individual who is missing and is being searched, the likelihood ratio may be computed using the following Bayesian network:

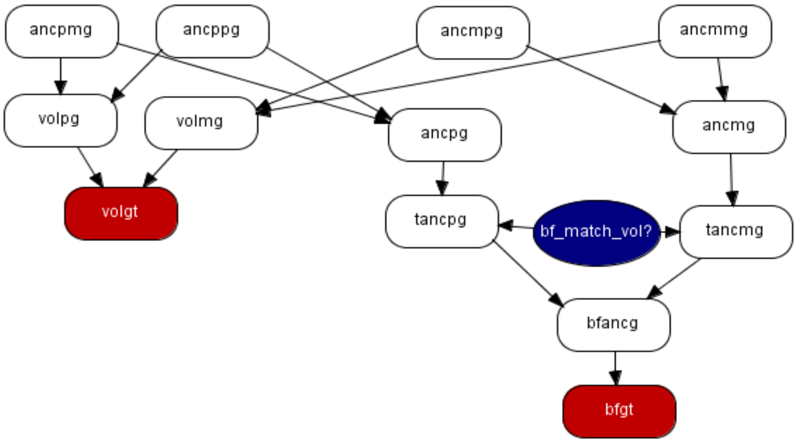

FIGURE 3: for civil identification with one volunteer, brother or sister relationship case.

Nodes **ancpmg** and **ancppg** are all of class founder and represent the volunteer's ancient genes of paternal and maternal inheritance. That inheritance will pass to nodes **volpg** and **volmg**, which are going to form volunteer genotype. Node **volgt** and **bfgt** are of class genotype, the volunteer and body found genotypes. The remaining nodes are the same as the ones presented in the previous problem.

5. CONCLUSION & FUTURE WORK

To connect an individual with a crime on the basis of a profile match may be dangerous because the database may contain undetected errors. In order to avoid misclassification with DNA from the database it is important to admit, at least, a second and independent analysis.

After computing the likelihood, whether it is a criminal case or a civil identification case, it is possible to compute the posterior odds, i.e., multiplying the likelihood ratio and the prior odds, in order to perform a comparative evaluation between the prosecution and the defense hypotheses.

The database file α is a subset of the population set P, $\alpha \subset P$. If the size of the database file is small, then one may only have a small fraction of the possible offenders. Therefore, it is important to take that into account. This topic should be considered in future work.

Whether it is criminal or civil identification in many situations the evidence may have more than one individual involved. In future work that must be considered. Also, in civil identification problems an important issue is to study how to compute the likelihood ratios when there is a match or a partial match between the genetic characteristic of the individual whose body was found and the file of "problem samples" and "reference samples", γ.

6. REFERENCES

1. A. P. Dawid, J. Mortera, V. L. Pascali, D. W. Van Boxel. *"Probabilistic expert systems for forensic inference from genetic markers"*. Scandinavian Journal of Statistics, 29:577-595, 2002

2. B. P. Battula, K. Rani , S. Prasad , T. Sudha. *"Techniques in Computer Forensics: A Recovery Perspective"*. International Journal of Security, Volume 3, Issue 2:27-35, 2009

3. David J. Balding. *"The DNA database controversy"*. Biometrics, 58(1):241-244, 2002

4. F. Corte-Real. *"Forensic DNA databases"*. Forensic Science International, 146s:s143-s144, 2004

5. G. Skinner. *"Multi-Dimensional Privacy Protection for Digital Collaborations"*. International Journal of Security, Volume 1, Issue 1:22-31, 2007

6. I. Evett and B. S. Weir. *"Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists"*, Sinauer Associates, Inc. (1998)

7. M. Andrade, M. A. M. Ferreira. "*Bayesian networks in forensic identification problems*". Journal of Applied Mathematics. Volume 2, number 3, 13-30, 2009

8. M. Andrade, M. A. M. Ferreira, J. A. Filipe. *"Evidence evaluation in DNA mixture traces"*. Journal of Mathematics and Allied Fields (Scientific Journals International-Published online). Volume 2, issue 2, 2008

9. M. Andrade, M. A. M. Ferreira, J. A. Filipe., M. Coelho. *"Paternity dispute: is it important to be conservative?"*. Aplimat – Journal of Applied Mathematics. Volume 1, number 2, 2008

10. M. Guillén, M. V. Lareu, C. Pestoni, A. Salas and A. Carrecedo. *"Ethical-legal problems of DNA databases in criminal investigation"*. Journal of Medical Ethics, 26:266-271, 2000

11. M. N. Anyanwu, S. Shiva. *"Comparative Analysis of Serial Decision Tree Classification Algorithms"*. International Journal of Computer Science and Security, Volume 3, Issue 3:230-240, 2009

12. P. Martin. *"National DNA databases – practice and practability. A forum for discussion"*. In International Congress Series 1261, 1-8

13. R. E. Neapolitan. *"Learning Bayesian networks"* , Pearson Prentice Hall, (2004)

14. R. G. Cowell, A. P. Dawid, S. L. Lauritzen, D. J. Spiegelhalter. *"Probabilistic Expert Systems"*, Springer, New York, (1999).