Report for the PIMS Hot Topics Workshop on Computational Criminology, September 19-21, 2012

November 12, 2012

1 Introduction and Synopsis of Conference Activities

The PIMS Hot Topics Workshop on Computational Criminology was held at IRMACS at Simon Fraser University in Burnaby, BC, Canada, on 19–21 September 2012. The goal of this workshop was to bring together mathematicians with an interest in working on crime modeling and analysis with researchers directly connected with real-world problems. This was a workshop based on computational criminology, an emerging field that takes the growing need for improved ways to use mathematics and computational techniques in understanding crime patterns and in developing methods for predicting and forecasting crime. The workshop included mathematicians and a group of PhD students and Post-Docs interested in the field as the core group, together with a small number of theoretical criminologists working in the field. The location of ICURS (Institute for Canadian Urban Research Studies) on the SFU campus with their expertise and database on urban criminology was an additional attractive feature of this event.

There were 47 registered participants from across Canada, the United States, Chile, Italy and the UK. Funding for the meeting was provided by PIMS, CEAMOS (Chile), IRMACS and the US Army RDECOM. The organizers were Alejandro Adem (PIMS, UBC), Andrea Bertozzi (UCLA), Patricia Brantingham (ICURS, SFU), Raul Manasevich (University of Chile) and Martin Short (UCLA). The conference featured ten plenary lectures by distinguished researchers in criminology and mathematical sciences:

Andrea Bertozzi (UCLA)
George Mohler (Santa Clara)
Gunnar Carlsson (Stanford)
Mario Primicerio (Firenze, Italy)
Jeff Brantingham (UCLA)
Raul Manasevich (CEAMOS,U.Chile)
Michael Ward (UBC)

Donald Brown (U.Virginia)
Milind Tambe (USC)
Theodore Kolokolnikov (Dalhousie)

In addition there were six lectures delivered in parallel sessions by

Maria D'Orsogna (Cal State Northridge)
Yves van Gennip (UCLA)
Jaime Ortega (U.Chile)
Richard Weber (CEAMOS, U.Chile)
Nancy Rodrguez (Stanford)
Rolando de la Cruz (PUC, Chile)

The PIMS website www.pims.math.ca/files/Hot_Topics_Abstracts_5.pdf contains abstracts for all the workshop presentations, and the plenary lectures are all freely available on the PIMS multimedia website www.mathtube.org

The workshop also featured breakout sessions to discuss future developments in the field. These provided a unique opportunity for researchers to compare their approaches to problems in criminology and lay the groundwork for growing international collaborations in this field of central relevance in society. The main themes discussed were *Data analysis in criminology*, *PDE models for crime hotspots* and *Game theory and Agent-based models*.

Below we summarize the discussions for the session on *Data Analysis and Criminology*, which was the most important one from a scientific point of view and led to very fruitful discussions. The summary was prepared by Yves Van Gennip.

2 Breakout session on Data Analysis and Criminology

This session brought together scientists from diverse disciplines such as criminology, mathematics, anthropology, and computer science, to exchange ideas on open problems in data analysis for criminology and the mathematical challenges that these pose.

2.1 Criminological questions

The general tendency during the breakout session, can be formulated concisely by two statements that were made at the start of the session: the mathematicians want more data to test their methods on, while the criminologists have a lot of data that needs analyzing. Here follow some examples of criminological questions that need addressing.

- A lot of criminological data is very high dimensional. For example, there is a large Vancouver data set that contains many details concerning all the crime reports over many years. Reducing the dimensionality of the data, by determining which details

are relevant to the questions at hand and which aren't, is a necessary step in making data sets manageable.

- Data sets often contain spatial or temporal errors. An example of the former occurs when the location of a crime is recorded as the police station where the record was made, instead of the actual location where the crime took place. A medical example of a temporal error can occur in the case of lead or mercury poisoning, when the symptoms become apparent only a long time after the poisoning has occurred. In order to draw justified conclusions from data, such errors need to be removed. Since the nature of these errors, and the extra information which could possibly be used to correct them, depends highly on the problem under consideration, it is a challenge to devise a general strategy to deal with them.

- In order to compare crime in different cities, with different layouts and road networks, and to make results from one city portable to other cities, a type of city-registration could be useful, comparable to, for example, brain registration in medical imaging. Network topology plays an important role here.

- Co–offender networks contain information on people who committed crimes together. Based on their social profiles it may be possible to predict missing links in the network, either uncovering unknown connections for past crimes, or predicting future criminal collaborations.

- Gangs often have an established core or top of their hierarchy that doesn't commit crimes anymore and that lets the people on the fringes do all the dirty work. Hence the people at the top often do not show up in data sets. Is there a way to deal with this, and other kinds of, incompleteness and imperfection of data?

- Geography is a huge influence on many sorts of crime. This includes both the actual physical geography creating boundaries in the landscape (rivers, hills, highways, ...), but also infrastructure connecting areas that are not physically close (again highways, railroads, subways, ...). One example of this is the formation of gang territories. Since conflict usually arises at the boundaries of territories, especially where one territory directly borders another one, it is important to understand the shape and formation of these.

- What is the life span of a gang? How do they get born (this may be connected to the emergence of hot spots in crime models) and which characteristics determine how long they survive? It was suggested that the severity of their entrance and membership rites is positively correlated with their life span, as is suspected for religions as well.

- When studying gangs, their fluidity has to be taken into account. People come and leave. Gang members interact with different gangs. Gang members victimize other

members of their own gang and members of other gangs (but possibly with different rates).

- There seems to be a correlation between landmarks and crime. For example, at an intersection with a store or a gas station, where cars are forced to slow down, crime rates are higher.

2.2 Mathematical challenges and problems

- Partial differential equations (PDE) can be used to model patterns. There are models out there for city growth and urban sprawl, as well as for formation of crime hotspots on simple domains. A coupling of these two kinds of models, where the urban sprawl model acts as a free boundary problem generating the domain for the crime hotspot model, could shed more light on the role the shape and development of cities plays on the formation of crime hotspots, for example through agglomeration. Transportation infrastructure also needs to be taken into account. For example, the distance between subway stops can, in a very real way, be considered shorter than the distance between a couple of blocks that needs to be traveled by foot.

- One way of relating such models (or any model) to reality, is by deriving scaling laws from them and corroborating them with empirical evidence. The PDE pattern formation literature knows many such examples, for example the study of coarsening rates.

- As already mentioned in the previous section, (large) spatial or temporal errors in recorded data can occur for a variety of reasons. An example of biased sampling coming from the reporting behavior, is the reporting of burglaries. Sometimes a burglary happens while the residents are at home, and so the temporal precision in the crime record is very high. However, a burglary that happened while the residents were away on several weeks of vacation, can usually not be pinpointed to a specific day or time with any accuracy. Regular biases need to be detected first and then corrected for.

- Different sets of data, all relevant for a particular problem or question, may be available in different resolutions. Mathematical techniques are needed to combine such data sets.

- When doing a point pattern analysis to find clusters of events, the real quantity that needs to be looked at is the ratio between the number of events and susceptible targets. Also uncertainty errors when fusing different layers of data, need to be represented. For example, if one needs to know where African-American elderly women live, but only data on African-Amercians, women, and the elderly, separately

is available, this data needs to be combined, taking into account different resolutions and uncertainty errors in each data set.

- A fundamental question is "what is a hotspot?" It is not clear how to define what a hotspot is, or if a single definition is even possible. For example, if someone calls the police ten times a week, does that mean the caller's location is a crime hotspot, or the caller just has a low tolerance for non-criminal events in his environment? In this case perhaps one can differentiate between calls and investigations? On the other hand, calls can come from both victims and offenders, so information might be lost as well.

  Subquestions that need to be tackled: "what is an event?", "what is the population (people, property, targets, ...)?" There are spatial and temporal relations to take into account. Populations change during the day (work population and entertainment population). Some crime is a byproduct of people, other of attractiveness. Behavior leading to crime is culture dependent, and models need to reflect this. In some groups the connectivity needs to reach a threshold for crime hotspots to appear.

  The fundamental problem of defining what a hotspot is, seems related to fundamental questions in other fields: "what is a pattern?" in pattern formation, "what is a good reconstructed image?" in image analysis, "what is a cluster?" in clustering analysis.

- If the concept of hotspot is (better) understood, also its dynamics should be studied.

- Crime hotspots can be very singular. Often there is no continuous transition between two hotspots. For example, in a region with many pubs, two of them can have many instances of assaults happening, while the others in between do not. This asks for mathematical models that can handle sharp edges, like total variation models.

- For high dimensional data, models and clustering on networks and hyper graphs is an interesting option. An example of this is the multiplex modularity optimization method in community detection. Also algebraic topology on simplicial complexes can be an interesting tool to study high dimensional data.

- In analogy with the field of image analysis, it would be very good for computational and mathematical criminology to have a set of standard bench mark 'ground truth' data sets available on which everybody tests their methods and algorithms. In this way it becomes easier to compare different methods, especially in the absence of clean definitions and goal (see the "what is a hotspot?" problem above). Artificial data, if it can be constructed, would be cleaner, but would lack real criminological relevance, while real data would be relevant, but be imperfect. It was suggested to have three or four different benchmark data sets, such that a wide range of possible questions can be asked about them. Suggestions that were made for candidates were data sets from Waterloo, Vancouver, and Los Angeles.

An example to compare this with, is the *Global Terrorism Database*[1] from the University of Maryland, which was initially built on data from the RAND Corporation (among other data) and is now the global standard for terrorism data.

3 Summary and Highlights

The meeting was extremely successful, as it brought together top researchers across several disciplines and jurisdictions to present state-of-the-art results in computational criminology. A notable aspect was the participation of Gunnar Carlsson from Stanford University, who is a pioneer in the applications of theoretical algebraic topology to data analysis. The involvement of his research team promises to add a new dimension to computational criminology. Plans are being laid for future meetings and long-term collaborations involving researchers in the US, Canada and Chile. One aspect under consideration is to expand this to a network in all of the Americas.

We also note that there were many young people at the meeting and we were also pleased to see a strong level of participation by women. The facilities at IRMACS were excellent. The organizers are grateful to PIMS for flawlessly organizing this event in such a short period of time.

[1] http://www.start.umd.edu/gtd