Crime Science
a SpringerOpen Journal

RESEARCH Open Access

Online social sports networks as crime facilitators

Bas Stottelaar, Jeroen Senden and Lorena Montoya*

Abstract

Emerging technologies such as broadband services and mobile and wireless technologies create not only benefits for the community but also risks (Choo, Smith, McCusker 78: iii, 2007). The implications of these developments should be evaluated to make any necessary changes to policing, policy and legislation. This study investigates the risk of disclosure of confidential information via online public exercise routes. The study identified in particular whether a) people inadvertently disclose their home address more often indirectly via online sports tracking networks than directly via other means and whether b) gender and age play a role in this disclosure. In addition, an analysis of the temporal characteristics of runs was performed to establish the window of opportunity for a home burglary and whether running is temporally predictable by hour of day or day of week. A total of 513 RunKeeper users were selected from the Dutch cities of Enschede and Nijmegen. 231 runners (45.03%) were located via RunKeeper and 122 (23.78%) via other Internet (i.e. non-social sports network) sources. It was found that a statistical difference exists between the indirect and direct disclosure of addresses; more runners disclose their home address via online sports tracking networks than via other sources. Furthermore, it was found that age played a role in the direct disclosure of addresses but not in the indirect disclosure. Older users more often disclosed their home address directly than younger ones. Conversely, gender plays a role in the indirect disclosure but not in the direct disclosure. Men more often disclosed their home address indirectly than women. Regarding temporal characteristics, it was found that the window of opportunity for a burglary is approximately 1 hour. Furthermore, the 'within subject' analysis suggests that the starting hour of the run is the most predictable temporal characteristic, followed by the duration of the run and the day of the week. This research ultimately shows the extent to which the unique combination of spatial and temporal information available in online sports tracking networks can enable criminals to predict where a potential target lives and when he or she will be out running.

Keywords: Leisure; Online social sports networks; Situational crime prevention; Spatio-temporal analysis; Home address location

Background

According to the 'routine activity approach' (Cohen and Felson 1979), three elements must converge in space and in time for crime to take place: a) a suitable target (person or product), b) a likely offender and c) absence of a capable guardian. People can facilitate their victimization by deliberately, negligently or unconsciously placing themselves at special risk even when they do not take an 'active' part in the crime (Sparks 1982). Hindelang et al. (1978)'s 'lifestyle-exposure theory of personal victimization' complements the previous views. This theory argues that victimization risk is a function of lifestyle, and in particular, that patterns of leisure expose people to victimization opportunities. Both the part of the population that reports spending leisure time online and the total time spent online have been increasing since 2008. However, the internet is changing leisure patterns since the total leisure time remains constant (Wallsten 2011). This study investigates whether the online publishing of running activities on sports social networks increases runners' vulnerability to crime in general, and home burglary in particular.

Protecting personal information is important to prevent victimization. Specific privacy concerns of online

*Correspondence: a.l.montoya@utwente.nl
Centre for Telematics and Information Technology, University of Twente, Hallenweg 19, 7522 NH, Enschede, The Netherlands

social networking include inadvertent disclosure of personal information, damaged reputation due to rumours and gossip, unwanted contact and harassment or stalking, surveillance-like structures due to backtracking functions (i.e. retracing actions), use of personal data by third-parties, hacking, and identity theft (Boyd and Ellison 2007). According to the Federal Bureau of Investigation (2014), predators, hackers, business competitors, and foreign state personnel troll social networking sites looking for information or people to target. News items and police websites often report that on-line sites such as Google Earth Street View, Facebook and Twitter are being used by burglars to target homes and businesses (e.g. Douglas County Sheriff 2013). Several sources have reported that many convicted burglars think that other burglars use social networks to identify targets (e.g. Distinctive Doors 2013). Moreover, a survey of 69 former burglars indicated that checking social media status is a favourite way of identifying target burglary homes (Edith Cowan University 2011, McMillan 2012). This seems to indicate that it is no longer the case that burglars are opportunistic and operate only on-the-ground, but that there are also technologically-savvy ones who operate in a more premeditative manner. In addition, it indicates that social media is used for the planning of a wide range of crimes.

Several studies (Ibrahim 2008, Tufekci 2008, Waters and Ackerman 2011) found that online social network users constantly balance perceived privacy risks and expected benefits. The most important benefit of online networks is probably the social capital resulting from creating and maintaining interpersonal relationships and friendship (Ellison et al. 2007). Studies reveal a 'privacy paradox' which is the disparity between reported privacy attitudes and observed privacy behaviours. In a study of online social network use and privacy, for example, those with Facebook profiles had greater concerns about strangers obtaining personal information about them than those who didn't have such profiles. However, among those with profiles, there was no relationship between participants' privacy concerns and the likelihood of them providing this information on the website (Stutzman and Kramer-Duffield 2010).

The research by (Madden and Smith 2010) and (Kramer-Duffield 2010) reveals four important general issues regarding personal information found online:

1. While basic contact information continues to top search lists, demand for social networking profiles and photos has grown considerably over time.
2. Young adults (i.e. ages 18-29) more often limit the amount of online information available about them than older adults.
3. Internet users are now more likely to search for social networking profiles than they are to search for information about someone's professional accomplishments or interests.
4. Females are more likely to have a friends-only Facebook account than males.

With regards to address information, Acquisti and Gross (2006) indicated that 24% of the Facebook users disclose their home address. Madden and Smith (2010) reached a similar conclusion (i.e. personal address disclosure of 26%) and also found that 23% of the users were unaware whether their home address could be found online. The research by boyd and Hargittai (2010) shows that people change their privacy settings more often than before. However, the authors consider that further research is needed to establish whether people fully understand the effect of the privacy setting changes they chose. These figures highlight that emerging technologies create not only benefits but also risks (Choo et al. 2007) and in particular, that information and communication technologies (ICT) lead to new crime opportunities. The Dutch Police (2012), for example, advises people to use social networks with caution and warns against sharing holiday information, photos or their current location. Although it is not wise to disclose personal information such as home address on the Internet, the figures show that this is common practice. This research answers the question of whether people unknowingly disclose address information indirectly more often using online social sport networks than directly via other online sources.

The rapid development of mobile applications has driven smartphone adoption. An example of a new type of smartphone-based mobile application is online sports tracking. This type of application is able to record data about exercises such as lifting counts during body building or recording the cycling or running speed on a map. The underlying technology behind the latter example is called geo-location. Geo-location uses data acquired from a radio or network connection enabled device to identify or describe the actual physical location. Despite its many benefits, geo-location does increase risk (ISACA 2011). Furthermore, when tracking allows somebody else's location to be traced, it is a sensitive issue from a privacy point of view (Klerks and Kop 2008). The risk, security, privacy and ethical concerns of geo-location and tracking are most often discussed in the context of enterprises and less often in the context of leisure. Leisure is important because people in many countries nowadays have more time available for it (Aguiar and Hurst 2007, Klerks and Kop 2008, OECD 2009).

Examples of current sport applications include RunKeeper, Endemondo, Strava, MapMyFitness, Nike+, Zombies, Run! and SportsTracker. Such applications are largely used by runners and have created a social network where people store recorded workouts and share their favourite

routes with others. This type of software can be download by anyone and installed on a wide range of computer systems (e.g. smartphones, tablets, laptops, desktops) but the GPS-enabled smartphone plays a key role since the route data is typically collected and uploaded onto the internet by means of this device. The user either creates a new account or connects using e.g. Facebook login information. Invitations for joining one's network are sent out in a very similar manner as in Facebook or LinkedIn. This step is optional because the user can choose not to send invitations, which still gives him access to all the public routes of other runners. In the particular case of RunKeeper, and at the time of writing this, it was possible (but not compulsory) to list one's location (e.g. city, country), the type of sport activity (e.g. running, swimming, cycling, skiing) and add a profile picture. The user could specify, based on the type of person (i.e. everyone, friends or nobody), the type of data (e.g. activities, activity maps, fitness reports, background activities, general body measurements) to be shared. It was also possible to prevent the user profile from showing up in search results. Another interesting feature from the security and privacy viewpoints was the option to connect to other sport or health-related applications.

Sports tracking applications introduce a new type of problem because the sharing of routes implies the disclosure of information about personal routine activities. The data available enables the identification not only of the route the runner followed but also of its temporal characteristics (i.e. when the run took place and its total duration). Runners are therefore disclosing the temporal pattern of their sport activities but because most of them start and stop their run at home, they also unwarily reveal their home address.

Burglary can result from personal information disclosure. The disclosure of a runner's home address as well as the temporal pattern of the sports activity facilitates crime. The relationship status on a social network such as Facebook or photos posted on Facebook or RunKeeper could be used to assess the likelihood that the person lives alone or accompanied. This is relevant because burglars prefer to avoid occupied homes (Bennett 1992, Bennett and Wright 1984, Hakim et al. 2001, Rengert and Wasilchick 2000, Waller and Okihiro 1978, Winchester and Jackson 1982) which explains why the most vulnerable homes are those of single-persons, single-parents, and younger-occupants (Rengert and Wasilchick 2000, Winchester and Jackson 1982). Although the duration of the absence from home strongly predicts burglary risk (Weisel 2002), many sources indicate that it takes burglars less than 10 minutes to break into a house and leave with the stolen items (e.g. Cusson 1993, Safewise 2013). In the case of running, the capable guardian is likely to be away and the suitable target objects are most likely at home since mobile phones are among the few items carried during running. Clarke (1999) discussed 'hot products' as items that attract attention and are targeted by thieves. Offenders focus on relatively few 'hot products', such as cars, laptop computers, DVD players, and mobile phones (Clarke and Eck 2005). The Internet and social networks in particular, provide a platform for identifying 'hot products'. The Dutch Police (2012) states that 80% of burglary is opportunistic ('thieves have Facebook too'). For example, posting a photo of one's new ultrabook computer and of the sports data via social media can make the runner's house a suitable target. In contrast, posting a photograph of one's expensive bicycle and of the sports data via social media might make the cyclist a suitable target to be robbed while out cycling. Internet searches reveal that in some countries it is common for cyclists to be forced to hand over their bicycles to criminals whilst out cycling (Miller 2013; Sapa 2013).

In summary, the present paper argues that sensitive information such as a home address coupled with the spatio-temporal characteristic of a workout creates (or increases) opportunity and facilitates several forms of victimization. In addition, on the Internet an offender does not have to come face-to-face with a potential target, which might make the act of target victimization easier (Petee et al. 2010). Online social sport networks provide relevant information for what in the field of museum theft is known as 'silently planned crime', which is crime that despite having low probability, can be prepared over a long period of time and which, at least in its planning phase, entails low detection risk. Erez (1980) argues, for example, that what may appear to be a spur of the moment crime could have been in the mind of the offender all along. Figure 1 shows the steps involved to identify the home of a runner based on online social sport network data. Only one step involves an 'on the ground' activity.

To the best of our knowledge, no research has been conducted about the potential of sports tracking apps as crime facilitators. However, there are two relevant general studies involving geo-location-enabled mobile applications. Dillon-Scott (2011) concluded that the majority of respondents are concerned about sharing their location without consent (84%), having their personal information or identity stolen and suffering loss of privacy (83%). Similarly, GSMA (2013) found that 92% of respondents in their survey want to be asked for their permission before sharing their location with a service or an application. ISACA (2012) indicates, however, that a majority find that the risk and benefits of location-based applications and services are appropriately balanced, showing that although people are apprehensive about privacy-breaches, they are likely to share their location via sports tracking applications since they deem the risk involved to be acceptable.

This study aimed to answer the following research questions:

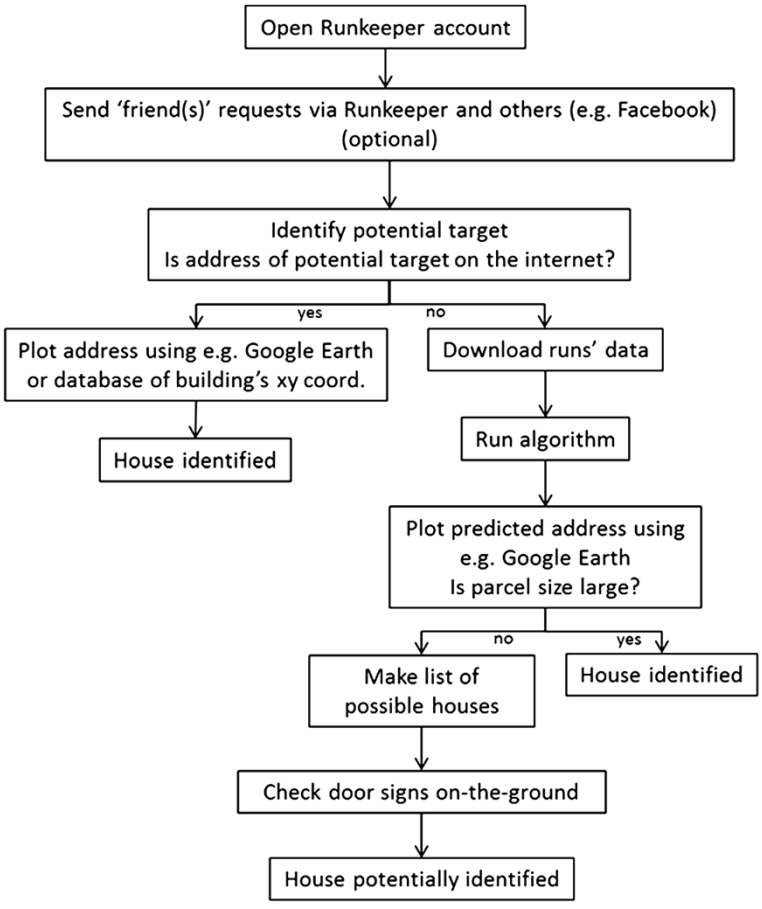

Figure 1 Steps involved for home location. The flowchart illustrates the sequence of steps necessary to identify a runner's home location.

1. What is the accuracy of an address identified on the basis of routes available on social sports networks and what is its implication in the Dutch urban context?
2. Are people more likely to disclose their address 'indirectly' (i.e. via running routes published on sports tracking networks) than 'directly' (i.e. by other sources such as Facebook, LinkedIn, Twitter, Yellow pages and company websites)?
3. Is there a relation between the runners' age and gender and the disclosure of an address?
4. What is the window of opportunity for a burglary?
5. Is running temporally predictable?

Although studies exist about the relation between crime and online leisure activities, most relate to dating sites, chat rooms and Facebook. In addition, in many other studies the concept of 'leisure' is operationalized as 'going out at night during weekends' (e.g. Gottfredson 1984). The present research is therefore concerned with a different form of leisure activity (i.e. running). The contribution of this study is the insight into the new phenomenon of online social sports tracking networks and in particular, its potential as a crime facilitator. The general approach of the research is to look into the routine activities of the target rather than of the offender. In addition, this research developed an algorithm to determine a home address based on public running routes published on online social sport tracking networks.

Methods

Sample

The aim was to select a random sample of runners who use an online sports social network to record their running route. The runner characteristics measured were gender and age. For the sample size selection, this study followed Bartlett et al. (2001)'s suggestion to use Cochran's equations[a].

Estimating the population of runners using sports tracking applications is difficult because the number of users per country is not listed in the websites. In addition, this research focuses on RunKeeper, which is a very popular application in The Netherlands, but there are other popular applications. According to Bottenburg (2006), in The Netherlands one in ten inhabitants run (refer

to the section 'Measures' for the definition of 'runner'). Given this figure and a combined population of Enschede, Nijmegen and the nearby cities Hengelo and Arnhem of approximately 550,000, the number of runners in this area was estimated at 55,000. Given an assumption of 1,000 RunKeeper users in these cities (which one might consider a conservative guess), the minimum sample size for the study is 250 runners. An assumption of 5,000 users would have yielded a sample size of 319 runners. Above 6,820 runners, the minimum sample size is 341 runners. It is worth noting that the cities of Hengelo and Arnhem fall within the search radius of Enschede and Nijmegen and are therefore taken into account in this population calculation.

To the best of our knowledge, there is no evidence that in The Netherlands running is geographically determined. Since running is easily accessible and location independent, it was assumed that the chosen cities were representative of The Netherlands as a whole.

The RunKeeper search engine was first queried to show runners that had recorded at least one route. A random sample was then drawn to select the runners for the study. The sample consisted of 513 runners with a total of 15,471 routes (i.e. approximately 30 routes per runner). The unit of analysis of the address disclosure evaluation is the individual runner whilst the route constitutes the unit of analysis of the temporal evaluation. For the temporal analysis, the sample size was 14,444 since some routes had no timestamp.

Measures

A **runner** is a person above the age of 6 who runs at least once a week (Bottenburg 2006). **Indirect address disclosure** occurs when an address is inferred (i.e. predicted) by means of exercise route data published in online social sports networks. **Direct address disclosure** occurs when an address is published in analogue or digital media such as the phone book or other similar directories. **Accuracy** defines how close a measurement is to the real value (i.e. the true home location). **Precision** is defined as how close together or how repeatable the results from a measurement are (SABS Standards Division 2012). Figure 2 depicts the possible combinations of accuracy and precision that can be used in a cluster analysis:

1. points are both precise and accurate,
2. points are precise but inaccurate,
3. points are imprecise but accurate and,
4. points are neither precise nor accurate.

The variables 'direct' and 'indirect disclosure of home address' are categorical variables (i.e. no-yes). Dichotomous variables are used to describe personal characteristics i.e. 'gender' (i.e. male vs. female) and 'age' (i.e. under 35 vs. 35 or older). The duration of the run is measured in minutes. Regarding distance measurements for the algorithm development, the (continuous) variables used were a) distance in kilometres between the start and end point of a route and b) number of GPS points per kilometre.

Procedure

The procedure to determine the direct and indirect disclosure of home address and personal information consisted of 3 basic steps.

1. Algorithm development, calibration and evaluation. The algorithm was developed to determine the home address of a runner based on his or her public running routes. The input for the algorithm is raw GPS data, grouped by runner. The algorithm classifies routes into suitable and unsuitable ones based on the quality of the GPS recording. In addition, it classifies the suitable routes into circular and non-circular. Spatial analysis of the start-finish points of circular routes was performed to identify the average point (i.e. predicted home location). The minimum distance thresholds and minimum routes per cluster were calibrated to yield the most reliable results (i.e. precision). If there are insufficient routes per cluster, non-circular routes having a starting point near the average point are added to the analysis. 18 volunteer runners participated in the evaluation of the algorithm and their 2012 routes were used to generate a map of their predicted home address. Subsequently, the runners were asked if this predicted location was accurate and to estimate the error. In other words, the first part of the analysis (i.e. cluster analysis) measures precision whilst the second part, involving

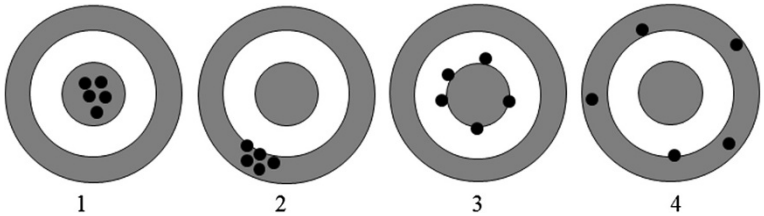

Figure 2 Accuracy and precision combinations. The diagrams depict four possible combinations of accuracy and precision for point data.

the volunteer runners, measures accuracy. Refer to Appendix 1 for further details on the algorithm.

2. Selection of runners and routes. The search webpage of RunKeeper was used to select runners from the cities of Enschede and Nijmegen. Routes of 3, 4, 5, 7.5 and 10 kilometres were searched since these constitute typical running distances. The search for routes yielded 513 unique random usernames of runners (see Figure 3 for an example of a map showing the result of the search for routes of one runner, with individual runs depicted with different colours).

   All public workouts were requested for the period between January 2011 and August 2013. This implies that the search narrowed down by selecting only those runners who had at least one recorded route in RunKeeper. The search engine did not search for specific routes, but for routes within a certain distance. Consequently, a search for a distance of 4 kilometres also yields routes of 3 kilometres. It is possible that a user may have added a route in Nijmegen while living in Amsterdam.
3. Home address and personal information search. The algorithm was applied to 513 runners to identify whether it succeeded or not in estimating a home address. To establish the age and gender of runners, the Runkeeper profile photos were assessed and a search was carried out by other means such as Facebook, Twitter, LinkedIn, the Yellow Pages and also company websites.

Statistical analysis

Three separate analyses were conducted for the address disclosure evaluation. The first one involved testing whether there were significant differences between direct and indirect disclosure of addresses. The ratio of indirect to direct disclosure was then calculated. The second one tested the relation between one type of disclosure (direct or indirect) and an independent variable (e.g. direct disclosure and age). For both the first and second analyses, cross tabulations and the Chi-square statistics were obtained. To make a straight-forward comparison between younger and older runners, the ratios of indirect to direct disclosure were computed. Similarly, the ratio of indirect to direct disclosure was calculated for females and males. The third analysis consisted of a multinomial logistic regression to model the disclosure of data based on both age and gender. The dependent variable consisted of 4 categories: a) no disclosure (used as reference category), b) indirect disclosure, c) direct disclosure and d) both indirect and direct. The output of the regression is a relative risk ratio (RRR) which is similar to the odds ratio used in logistic regression. The standard interpretation of the relative risk ratio is for a unit change in the independent variable, the relative risk ratio of the dependent variable *m* relative to the reference category is expected to change by a factor of the respective parameter estimate given that the variables in the model are held constant (Institute for Digital Research and Education 2014).

The temporal evaluation involved three types of analysis. The opportunity window of a burglary was obtained by calculating the mean duration of runs. To test whether individuals do runs of similar duration, a one-way analysis of variance was performed and the intra-class correlation (ICC) was computed[b]. Since the confidence intervals are computed under the assumption that *rho* is normally distributed, it is appropriate to extend the assumption for

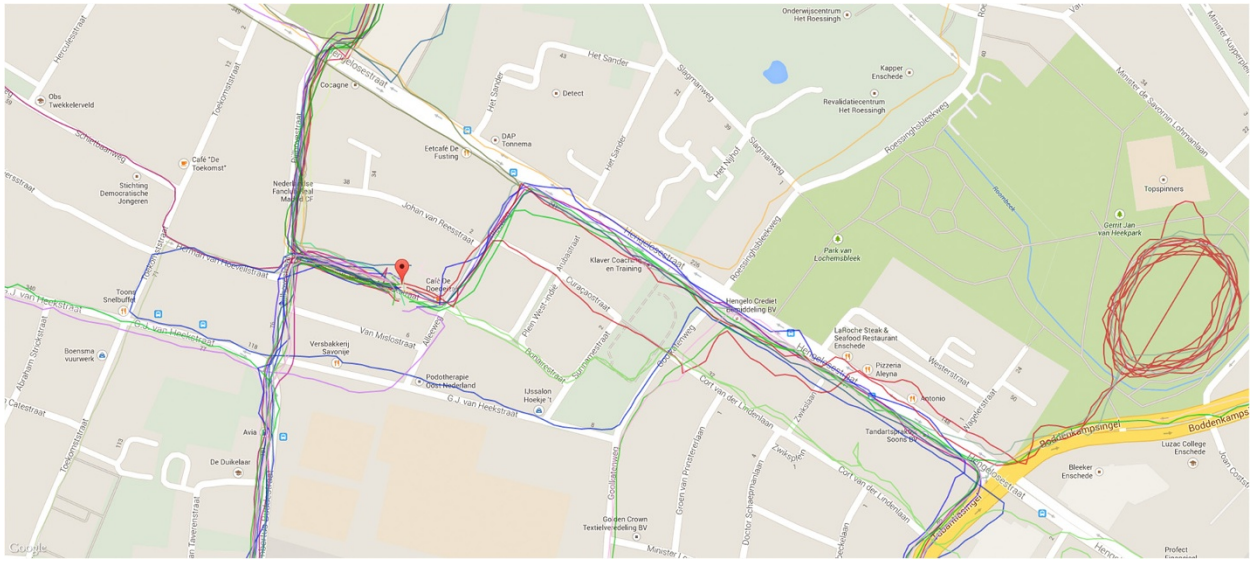

Figure 3 Example of route output. The map shows the result of the search for routes of one runner.

providing a simple method to test the difference between two ICCs. This method involves computing standardized scores (i.e. z-scores) using the following formula (Newson 2002):

(c) $z = \frac{(icc1 - icc2)}{\sqrt{(se1^2 + se2^2)}}$

Runners that recorded only one route where not taken into account for the calculation of the ICC.

Results

Disclosure of address

The validation work showed that the chosen algorithm thresholds are sufficient to yield accurate results. The average estimated error was 45 metres ($SD = 74,8$). It is worth mentioning that simple consumer type GPS receivers, like the ones that are incorporated into smartphones, are only capable of measuring positional accuracy to within a few tens of metres. For example, Garmin (2013) claims that their receivers are accurate to within 15 metres.

Table 1 shows that 53 runners were located only directly, 162 only indirectly, 69 via both methods and 229 could not be located. For 231 runners, the algorithm yielded 476 possible home addresses. This means that some runners have more than one home address, such as students living at the university and also with their parents. 66 out of 69 indirect addresses were matched very closely to the direct addresses. In three cases the direct home address was more than 200 metres away from the indirect home address. The results show a statistical significant difference between indirect and direct sharing of an address ($\chi^2 = 8.594$, $df = 1$, $p = 0.003$). The ratio of indirect to direct disclosure is 1.89.

The research tested the relation between the age of a person and the disclosure (i.e. indirect or direct) of the home location. Similarly, it tested the relation between the gender of a person and the disclosure of the home location. The results for age are presented in Table 2 and the results for gender in Table 3. Of the 513 chosen runners, age could not be determined for 130 of them. The results show a statistically significant relationship between the age of a runner and the direct disclosure of the address ($\chi^2 = 16.801$, $df = 4$, $p = 0.002$). However, there was no statistically significant relationship between the age and the indirect sharing of the address of a user. The ratio of indirect to direct disclosure is higher for younger people (i.e. 1.88) than for older people (i.e. 1.35). Gender could not be determined for 93 out of the 513 runners. There is strong evidence of a relationship between the gender and the indirect sharing of the address ($\chi^2 = 4.773$, $df = 1$, $p = 0.03$). There is no statistically significant relationship between the gender and the direct sharing of a home address. The ratio of indirect to direct disclosure of home addresses is higher for males (i.e. 1.74) than for females (i.e. 1.69).

Finally, the multinomial logistic regression model in Table 4 provides further information since so far, only the relation between one form of disclosure and one independent variable has been tested at a time (e.g. indirect disclosure with gender). The model shows that for females (relative to males), the relative risk for a) indirect address disclosure via Runkeeper (relative to no disclosure) would be expected to decrease by a factor of 0.44 ($p = 0.004$), b) direct disclosure (relative to no disclosure) would be expected to decrease by a factor of 0.48 (p = 0.055) and both direct and indirect disclosure (relative to no disclosure) would be expected to decrease by a factor of 0.49 ($p = 0.047$), given that the other variables in the model are held constant. Similarly, the model shows that for older runners (relative to younger ones), the relative risk for a) indirect disclosure via Runkeeper (relative to no disclosure) would be expected to increase by a factor of 1.02 (result not significant), b) direct disclosure (relative to no disclosure) would be expected to increase by a factor of

Table 1 Direct and indirect address search results

Method	Freq.	%
Direct	53	10.33
Indirect	162	31.58
Both Direct & Indirect	69	13.45
None (i.e. no disclosure)	229	44.64
Total	513	100.00

The table shows the number of RunKeeper runners for whom an address was found directly and indirectly.

Table 2 Direct and indirect address search results based on age characteristics

Age	Direct		Indirect		Total
	Freq.	%	Freq.	%	Freq.
Younger than 35	39	21.20	91	49.46	184
35 or older	80	40.20	107	53.77	199
Unknown	3	2.31	33	25.38	130
Total	122	23.78	231	45.03	513

The table shows the number of RunKeeper runners for whom an address was found directly and indirectly, according to age group.

Table 3 Direct and indirect address search results based on gender

Gender	Direct		Indirect		Total
	Freq.	%	Freq.	%	Freq.
Male	96	30.67	167	53.35	313
Female	26	24.30	44	41.12	107
Unknown	0	0.00	20	21.51	93
Total	122	23.78	231	45.03	513

The table shows the number of RunKeeper runners for whom an address was found directly and indirectly, according to gender.

Table 4 Multinomial logistic regression model

	RRR	SE		Confidence interval	
				Lower 5%	Upper 95%
RunKeeper					
Older	1.02	0.26		0.63	1.67
Female	0.44	0.13	**	0.25	0.77
Constant	1.23	0.23		0.86	1.77
Other online sources					
Older	2.03	0.68	*	1.05	3.91
Female	0.48	0.18	[a]	0.22	1.07
Constant	0.34	0.09	***	0.20	0.57
Both					
Older	3.04	0.99	**	1.61	5.76
Female	0.49	0.18	*	0.24	0.99
Constant	0.32	0.09	***	0.19	0.54

The model predicts the disclosure of home address information for RunKeeper runners.
'None' (i.e. no address disclosure) is the reference category.
*Significant: $p < 0.05$; **Significant: $p < 0.01$; ***Significant: $p < 0.001$.
N = 383 ($\chi^2 = 27.69$; $p<0.001$).
RRR: Relative risk ratio, SE: Standard error.
[a] $p = 0.055$.

2.03 ($p = 0.035$) and c) both direct and indirect disclosure (relative to no disclosure) would be expected to increase by a factor of 3.04 ($p = 0.001$), given that the other variables in the model are held constant.

Temporal characteristics

The window of opportunity for a burglary is on average 52.85 minutes (SD=57.58). The correlation of run duration within runners (i.e. ICC) is 0.13 ($F = 11.11$; $df = 210/14,213$; $p = 0.000$). Regarding gender differences, the correlation of run duration for female runners is 0.17 ($F = 11.14$; $df = 159/11,458$; $p = 0.000$) whilst it is 0.13 for males ($F = 2.39$; $df = 34/1,989$; $p = 0.000$). With regards to age differences, the correlation of run duration for younger runners is 0.15 ($F = 9.58$; $df = 130/6,768$; $p = 0.000$) whilst it is 0.11 for older ones ($F = 12.89$; $df = 55/5,365$; $p = 0.000$).

The correlation of hour of run start within runners is 0.20 ($F = 18.27$; $df = 210/14,213$; $p = 0.000$). Regarding gender differences, the correlation of hour of run start within female runners is 0.16 ($F = 11.35$; $df = 34/1,989$; $p = 0.000$) whilst it is 0.20 for males ($F = 19.31$; $df = 159/11,458$; $p = 0.000$). With regards to age differences, the correlation of the hour of run start for younger runners is 0.19 ($F = 12.31$; $df = 139/6,768$; $p = 0.000$) whilst it is 0.18 for older ones ($F = 21.75$; $df = 55/5,365$; $p = 0.000$).

The correlation of day of week within runners is 0.02 ($F = 2.16$; $df = 210/14,213$; $p = 0.000$). Regarding gender differences, the correlation of day of week for female runners is 0.01 ($F = 1.85$; $df = 34/1,989$; $p = 0.002$) whilst it is 0.02 for males ($F = 2.29$; $df = 159/11,458$; $p = 0.000$). With regards to age differences, the correlation of day of week for young runners is 0.02 ($F = 2.10$; $df = 139/6,768$; $p = 0.000$) whilst it is 0.01 for older ones ($F = 2.08$; $df = 55/5,365$; $p = 0.000$). Refer to Tables 5 and 6 for an overview of the ICC analysis.

Discussion

This research has shown that a runner's home location can be predicted via sports tracking application data. A home address, together with other information, can increase crime opportunity for burglars or identity fraudsters. An algorithm was described to determine a home address based on public workouts. For 36 out of 69 runners both directly and indirectly located, the home address matched.

The first research question related to the accuracy of an address identified on the basis of online social sports networks and its implication in the Dutch context. The estimated average error of the algorithm is 45 metres. Assuming a typical Dutch urban neighbourhood consisting of single family houses (i.e. parcel widths of 10 metres, houses on both sides of the road with back entrances), the algorithm narrows the estimated position down to within 8 possible houses. A quick field expedition to check family name signs on the front door of houses would enable the correct house to be identified. In addition, although evidence shows that burglars do not necessarily target more affluent areas (Bernasco and Nieuwbeerta 2005) (since these have often higher security), it is possible that those who go to the trouble of preparing a burglary using online sources might consider the trade-offs of targeting more affluent areas over less affluent ones. Since the parcel sizes in more affluent areas are generally larger, this algorithm would most likely yield the exact house or the one next door. The same would apply in other countries where the

Table 5 Intraclass correlations for gender

Variable and categories	ICC	SE	Confidence interval		Z
			Lower 5%	Upper 95%	
Run duration					
Male (ref. cat.)	0.13	0.02	0.08	0.17	
Female	0.17	0.05	0.01	0.02	0.74
Hour of run start					
Male (ref. cat.)	0.20	0.03	0.14	0.27	
Female	0.16	0.05	0.06	0.25	-0.80
Day of week					
Male (ref. cat.)	0.02	0.00	0.01	0.03	
Female	0.02	0.01	0.00	0.03	0.00

The table shows the correlations within runners, standard errors, confidence intervals of the ICC and z-score.
'Between subject' N = 134; 'within subject' N = 1,989; Z: z-score.

Table 6 Intraclass correlations for age

Variable and categories	ICC	SE	Confidence interval		Z
			Lower 5%	Upper 95%	
Run duration					
Younger than 35 (ref. cat)	0.15	0.03	0.09	0.21	
35 or older	0.11	0.03	0.05	0.17	0.71
Hour of run start					
Younger than 35 (ref. cat)	0.19	0.04	0.12	0.26	
35 or older	0.18	0.05	0.09	0.27	0.17
Day of week					
Younger than 35 (ref. cat)	0.02	0.01	0.01	0.03	
35 or older	0.01	0.00	0.00	0.02	1.02

The table shows the correlations within runners, standard errors, confidence intervals of the ICC and z-score.
'Between subject' N = 139; 'within subject' N = 6,768; Z: z-score.

parcel sizes are in general larger (e.g. USA). Although the runners that took part in the algorithm validation had 'private' workouts, most were surprised about the accuracy of the algorithm.

The second question aimed to identify whether a statistically significant difference existed between the indirect disclosure and the direct disclosure of a home addresses. Table 1 shows that runners tend to disclose their address indirectly more often than directly. The ratio of indirect to direct disclosure is 1.89. 23% of the participants of the research conducted by Madden and Smith (2010) did not know if their home address (direct address in this case) was available online. We suspect that a higher percentage might be unaware that it is possible to determine an address based on sports data.

The third research question regarded the relation between the age and gender of a runner and the disclosure of the home address (either directly or indirectly). The ratio of indirect to direct disclosure of home addresses is higher for younger people than for older people. A possible explanation is that older people are more likely to be found in the phone book, since younger people have not settled down yet and/or prefer a mobile telephone line instead of a fixed line. Disclosure of home address via both methods increases with age. The findings support the view of Madden and Smith (2010) who found that the disclosure of home address increases with age. Older people possibly have less understanding of Internet-based risks. Regarding gender, the ratio of indirect to direct disclosure of home addresses is higher for males than for females. No literature was found to explain this difference but we suspect that females might be more cautious, possibly for fear of assault. Women are less likely to disclose their home address via both methods than males (result is marginally significant). This finding is in line with Stutzman and Kramer-Duffield (2010), who found that women are more cautious than males with regards to online privacy.

The fourth research question aimed to identify the window of opportunity for a residential burglary based on the running activity. It was found that there is a window of approximately one-hour. Since many sources (e.g. Cusson 1993; Safewise 2013) indicate that most burglars spend less than 10 minutes at the crime scene, such a window provides sufficient time to carry out the crime.

The fifth research question related to whether previous temporal characteristics predict future temporal characteristics for the same runner. The general finding is that the hour of start of a run is the most predictable temporal variable, followed by its duration and day of week. The temporal predictability always decreases with age but the results are mixed for gender. Males have more predictable patterns with respect to the run's starting hour and the day of week whilst females are more predictable with respect to the duration. Knowledge of this temporal predictability would probably increase the confidence of the motivated criminal.

RunKeeper offers three types of privacy settings: public, friends only and private. This research used public workouts which can be viewed by anyone, not requiring a RunKeeper user account. On average, the 18 runners who participated in the algorithm validation provided 24 routes per user whereas the 513 runners provided 30 routes per user. For 66 runners the address matched both directly and indirectly despite the 45 metre error of the algorithm. However, the accuracy of home addresses and the number of located runners could be increased if friends-only workouts were to be used.

This research identifies previously undetected crime risks that could be easily reduced. Regarding recommendations to reduce risk, an awareness campaign about the risks involved would probably have an impact since the participants in the algorithm validation were surprised about its accuracy and hence unaware of what can be achieved by analysing running data. In particular, users of these types of networks should be made aware of their potential for data mining which refers to the automatic or semi-automatic data analysis to extract previously unknown patterns. In addition, since some runners sometimes share their 'favourite' routes with strangers because they wish to suggest interesting, demanding or simply 'nice' runs, it would be advisable to remove the starting and ending portions of the route. The software developer should highlight the risks of runners sharing their routes, particularly when they start and end at home. In addition, an automatic removal of the starting and ending portions of all routes could be performed directly by the software developer. Such an action would constitute a known standard setting that would credit the developer with having a security policy to protect users.

By preventing a potential burglar or robber from finding a temporal pattern in the running activity (i.e. routine

activity), his/her perception of risk would be increased, hence reducing the runner's likelihood of becoming the victim of a burglar, robber or of a predatory offender. This last measure constitutes situational crime prevention because it involves the modification of the (potential) crime settings, making criminal action less attractive to offenders.

Conclusions

The present research focused on runners but further research could be carried out to estimate the spatio-temporal pattern of cyclists, since bicycle theft and/or robbery while out cycling is a problem in some countries.

This research suggests that people might neglect the risks of inadvertently disclosing personal information that could increase their susceptibility to victimization possibly as the wish to socialize and to highlight their sporting achievements overrides their natural caution. Rather than looking into the routine activities of the criminal, this research has looked into the routine activities of the target. It shows how the unique combination of spatial and temporal information available in online sports tracking networks can enable criminals to predict with high likelihood where a potential target lives and when he or she will be out running. On the basis of these findings one can conclude that online sports tracking networks have potential to be part of the modus operandi of several types of crime, both at home and en-route. This research therefore shows that there is scope for traditional crime to become increasingly 'digitalized'.

Endnotes

[a](a) $n_0 = \frac{t^2 * (p)(1-p)}{d^2}$ (b) $n_1 = \frac{n_0}{1+{}^{n_0}/_N}$ where t is the value for the selected alpha level of 0.25 in each tail (i.e. 1.96), $(p)(1-p)$ is the estimate of variance (by manual inspection of workouts, we estimated $p = {}^1/_3$), d is the acceptable margin of error for the proportion being estimated (i.e. 5% or 0.05) and N is the population size. Formula (b) should be applied when n_0 exceeds 5% of the population.

[b]The ICC is a measure of agreement unlike the commonly used Pearson correlation which is a measure of association. Therefore, the standards that apply to a Pearson correlation do not apply to an ICC. For example, while a Pearson correlation of 0.3 may be considered small, an ICC of 0.3 is quite large.

Appendix 1: Algorithm

The first phase is pre-processing and is based on several properties that are calculated for each route. The following are the steps in the pre-processing phase:

1. For each route, the distance between the start point and end point is calculated.
2. For each route, the number of GPS points per distance unit is calculated.

The first metric allows non-circular routes to be filtered. The logic behind this is the following: running is an exercise that can be done everywhere but it tends to start and end at home; however, some people stop recording their workout just before they arrive home since they want to 'cool down'. Since this interferes with the statistics of the run, it is common that runners do not record the last part (i.e. the 'cool down' phase) of their run. Therefore, the algorithm takes a distance D between the start and the end into account.

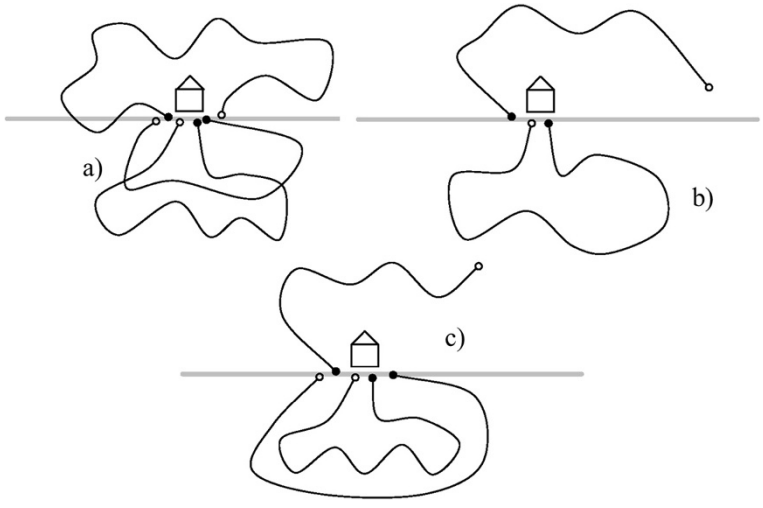

Figure 4 Basic examples demonstrating the algorithm. a) Start-stop of all routes close to each other **b)** Start-stop of one route exceeds a certain threshold and **c)** Invalid route supports the cluster of two valid ones due to its start position.

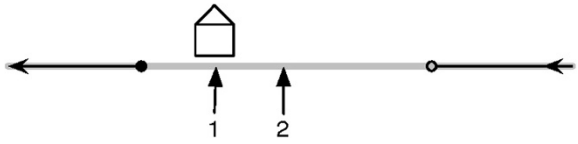

Figure 5 Weighted average point. Arrow 1 denotes the weighted average point for $W = 0.75$ and arrow 2 for $W = 0.5$ (normal).

The second metric allows the algorithm to distinguish between good and bad recordings. It is common that a runner sets a GPS to record at once a second (i.e. 1 Hz). Under optimal circumstances and a running speed of 12 kilometres per hour, a location is recorded every 3.33 metres. However, poor reception perhaps due to atmospheric factors, terrain, tree canopy or tall buildings can produce unsuitable recordings. A limit P was chosen as a lower bound on the number of GPS recordings per distance of 1 kilometre (PPD). This classifies recorded workouts (including manually entered ones) into suitable and unsuitable ones. All routes that comply with the chosen thresholds are referred to as 'valid routes' (i.e. candidate routes).

The third step involves the spatial analysis of clusters. Since the location of the houses is unknown, the metric used was precision. This step involves first iterating each valid route for each runner to determine the average start-stop point. Each average point was then compared to other average points and was allocated to a cluster if the distance between these two points was within range R (i.e. the average point delta distance). This implies that the points are precise. If a cluster of average points is above the threshold S, it is assumed to be a home address. If the cluster of average points is too small (i.e. it contains very few points), but is near the threshold S, all other routes (including invalid ones) outside the cluster are inspected for a starting point near the average point of the cluster which minimizes the distance. In this way, workouts that finished further away can still support the cluster if the starting point is close to the average start-stop. Figure 4 describes this reasoning.

1. Start-stop of all routes close to each other.
2. Start-stop of one route that exceeds a certain threshold.
3. Invalid route supports the cluster of two valid ones due to its start position.

The average point requires further refinement for two reasons. As already mentioned, often people finish their workout recording earlier to 'cool down' and walk home which means that the end point is not near the start point. Second, sometimes there is a delay in obtaining a GPS position and the initial readings may be inaccurate as the GPS acquires the satellite constellation. Less accurate home addresses are obtained by taking the average point as $P_{average} = (P_{start} + P_{stop})/2$. To tackle this problem, a weighted average is used and the average is defined as $P_{average} = X * P_{start} + (1 - X) * P_{stop}$. Figure 5 illustrates this issue.

Two additional figures were generated to verify the algorithm's thresholds. Figure 6 shows the number of routes per runner for all runners and for selected runners. An almost identical distribution of routes was selected compared to all processed routes. Since most runners have only a few routes available, a threshold of minimum cluster size $S = 3$ was considered reasonable. Figure 7 shows the frequency distributions of the route lengths, the delta distance and the points per distance for all the routes and for the selected routes. The first column shows that most routes are shorter than 20 km. The delta distance shows

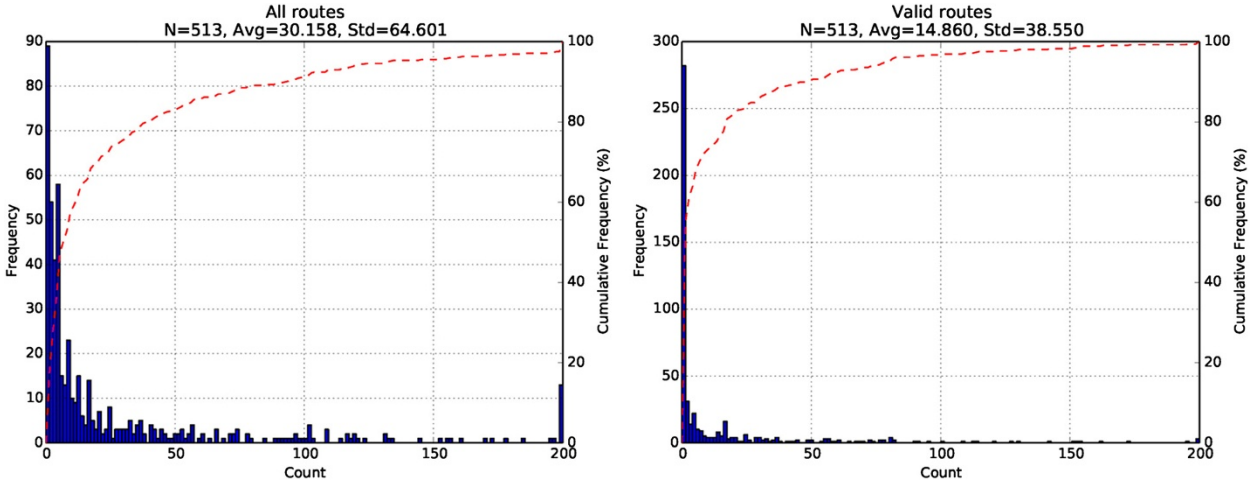

Figure 6 Number of routes per person. Distributions of the number of routes per person for all routes and for selected routes.

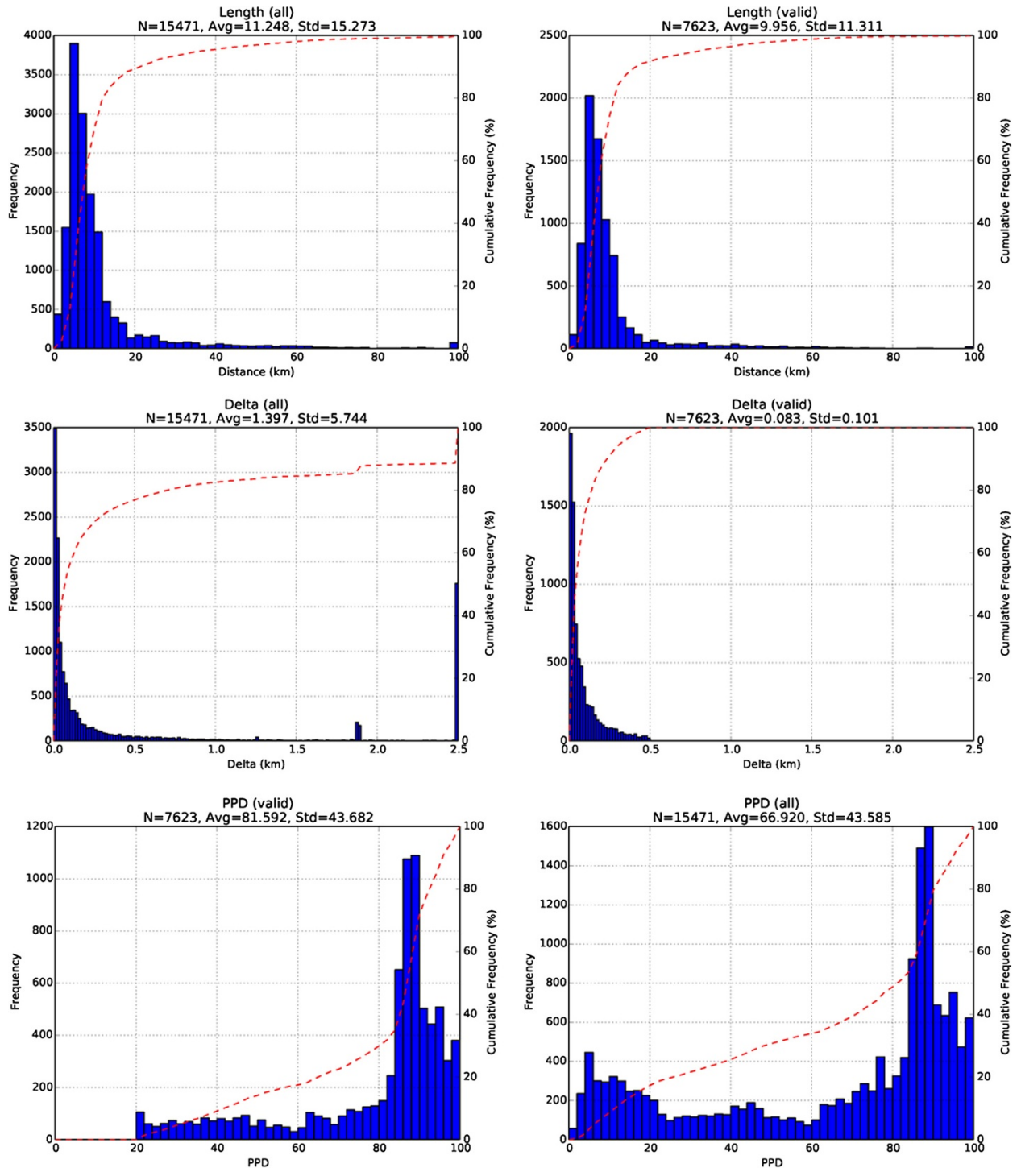

Figure 7 Route characteristics. Distributions of length, delta distance and points per distance for valid and for selected routes (in kilometres).

that in the context of running, the distance between start and end is usually below 500 metres. A delta distance of $D = 500$ metres was therefore chosen. A lower bound of $P = 20$ was chosen for the minimum number of points per distance (PPD). This implies a GPS recording is needed on average every 50 metres, which seems an adequate number to account for poor GPS signal reception, for up to 80 points per distance (PPD) of one kilometre. Figure 7 also shows that lowering the thresholds does not considerably affect the number of located runners, except for the

cluster size S. In this experiment, lowering the cluster size to $S = 2$ resulted in 287 located runners instead of 231, thus lowering its rigorousness.

An initial estimation and some experimenting with the thresholds therefore resulted in the following values:

- Delta distance $D = 500$ metres.
- Lower bound of $P = 20$ points per distance.
- Average point weight $W = 0.90\%$.
- Average point delta distance $R = 150$ metres.
- Cluster size $S = 3$.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
BS and JS carried out the data mining, developed the algorithm and participated in the drafting of the manuscript. LM conducted the statistical analysis and drafted the manuscript. All authors read and approved the final manuscript.

Authors' information
Bas Stottelaar and Jeroen Senden followed the computer security track of the computer science master degree programme at the University of Twente. Lorena Montoya is senior researcher at the Services, Cyber-security and Safety group of the University of Twente.

Received: 12 October 2013 Accepted: 12 May 2014
Published online: 22 August 2014

References

Acquisti, A, & Gross, R (2006). Imagined communities: Awareness, information sharing, and privacy on the facebook. In G Danezis & P Golle (Eds.), *Privacy Enhancing Technologies, volume 4258 of Lecture Notes in Computer Science* (pp. 36–58). Cambridge, UK: Springer.

Aguiar, M, & Hurst, E (2007). Measuring trends in leisure: The allocation of time over five decades. *Quarterly Journal of Economics, 122*, 969–1006.

Bartlett, JI, Kotrlik, JW, Higgins, CC (2001). Organizational research: Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, *19*(1), 43.

Bennett, T (1992). *Burglars' Choice of Targets*. London: Routledge.

Bennett, T, & Wright, R (1984). *Burglars on Burglary - Prevention and the Offender*. Hampshire, UK: Gower.

Bernasco, W, & Nieuwbeerta, P (2005). How do residential burglars select target areas?: A new approach to the analysis of criminal location choice. *British Journal of Criminology*, *45*(3), 296–315.

Bottenburg, M (2006). *De Tweede Loopgolf: Over Groei en Omvang van de Loopsportmarkt en hoe de Knau Haar Marktaandeel Verder Kan Vergroten*. 's-Hertogenbosch, The Netherlands: W.J.H. Mulier Instituut.

Boyd d, & Ellison, N (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230.

Boyd d, & Hargittai, E (2010). Facebook privacy settings: Who cares. *First Monday*, *15*(8), 2.

Choo, K-KR, Smith, RG, McCusker, R (2007). *Future Directions in Technology-enabled Crime: 2007-09,* Vol. 78. Canberra, Australia: Australian Institute of Criminology.

Clarke, RV (1999). Hot products: Understanding, anticipating and reducing demand for stolen goods. Report 112, UK Home Office.

Clarke, RV, & Eck, JE (2005). *Crime Analysis for Problem Solvers in 60 Small Steps*. Washington DC: U.S. Department of Justice.

Cohen, LE, & Felson, M (1979). Social change and crime rate trends: a routine activity approach. *American Sociological Review*, *44*, 588–608.

Cusson, M (1993). *Situational Deterrence: Fear During the Criminal Event* Vol. 1, (pp. 55–68). Monsey, New York, USA: Criminal Justice Press.

Dillon-Scott, P (2011). *Geolocation Apps Causing New Privacy and Safety Fears for Smartphone Users*: Sociable Publishing Ltd. http://sociable.co/mobile/geolocation-apps-causing-new-privacy-and-safety-fears-for-smartphone-users/.

Distinctive Doors (2013). Infographic: How burglars are using social media. http://www.distinctivedoors.co.uk/news/51-infographic-how-burglars-are-using-social-media.

Douglas County, Sheriff (2013). Crime prevention/social media alert. http://www.dcsheriff.net/newsroom/crime-prevention-social-media-alert/.

Edith Cowan, University (2012). Inside the mind of a burglar. http://www.ecu.edu.au/news/media-releases/2012/12/inside-the-mind-of-a-burglar.

Ellison, NB, Steinfield, C, Lampe, C (2007). The benefits of facebook "friends:"ï social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, *12*(4), 1143–1168.

Erez, E (1980). Planning of crime and the criminal career: Official and hidden offenses. *Journal of Criminal Law and Criminology*, *71*(1), 73–76.

Federal Bureau of, Investigation (2014). Internet social networking sites. http://www.fbi.gov/about-us/investigate/counterintelligence/internet-social-networking-risks.

Garmin (2013). What is gps? http://www.garmin.com/aboutGPS/.

Gottfredson, MR (1984). *Victims of crime: The Dimensions of Risk*. London, UK: HM Stationery Office.

GSMA (2013). Gsma reveals fears over mobile privacy are holding back the growth of mobile apps in latin america. http://www.gsma.com/newsroom/gsma-reveals/.

Hakim, S, Rengert, GF, Shachmurove, Y (2001). Target search of burglars: A revised economic model. *Papers in Regional Science*, *80*(2), 121–137.

Hindelang, MJ, Gottfredson, MR, Garofalo, J (1978). *Victims of, Personal Crime: an Empirical Foundation for a Theory of Personal Victimization*. Cambridge, Massachusets, USA: Ballinger Publishing Co.

Ibrahim, Y (2008). The new risk communities: Social networking sites and risk. *International Journal of Media and Cultural Politics*, *4*(2), 245–253.

Institute for Digital Research and Education (2014). Stata annotated output: Multinomial logistic regression. http://www.ats.ucla.edu/stat/stata/output/stata_mlogit_output.htm.

ISACA (2011). Geolocation: Risk, issues and strategies. Rolling Meadows, IL USA. http://www.isaca.org/groups/professional-english/wireless/groupdocuments/geolocation_wp.pdf.

ISACA (2012). 2012 geolocation use and concerns survey USA. Rolling Meadows, IL USA. http://www.isaca.org/Knowledge-Center/Research/ResearchDeliverables/Pages/Geolocation-Risks-Issues-and-Strategies.aspx.

Klerks, P, & Kop, N (2008). *Societal Trends and Crime-relevant Factors: an Overview for the Dutch National Threat Assessment on Organized Crime 2008 - 2012*. Apeldoorn, The Netherlands: Politieacademie.

Madden, M, & Smith, A (2010). Reputation management and social media. http://www.pewinternet.org/2010/05/26/reputation-management-and-social-media/.

McMillan, G (2011). Burglars now using twitter, facebook against you. http://techland.time.com/2011/09/27/burglars-now-using-twitter-facebook-against-you/.

Miller, D (2013). Cyclists beware: Gang of thieves target riders of expensive bicycles. http://www.khou.com/news/local/Bikers-beware-Gangs-of-thieves-target-riders-of-expensive-bicycles-218917611.html.

Newson, R (2002). Comparing two iccs. http://www.stata.com/statalist/archive/2002-06/msg00246.html.

OECD (2009). *Special Focus: Measuring Leisure in OECD Countries*, (pp. 19–49). Paris, France: OECD Publishing.

Petee, TA, Corzine, J, Huff-Corzine, L, Clifford, J, Weaver, G (2010). Defining "cyber-crime": Issues in determining the nature and scope of computer-related offenses. In T Finnie, T Petee, J Jarvis (Eds.), *Futures Working Group*, volume 5 (pp. 6–11). Quantico, VA, USA. PFI/FBI.

Rengert, GF, & Wasilchick, J (2000). *Suburban Burglary: A Tale of 2 Suburbs*, (2nd. ed.) Springfield, Illinois, USA: Charles C Thomas Pub Ltd.

JCGM (2012). International vocabulary of metrology: Basic and general concepts and associated terms (vim). http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf.

Safewise (2013). You just got robbed (and it only took 10 minutes). http://www.safewise.com/you-got-robbed.

Sapa (2013). Three held for robbing cyclists. http://www.news24.com/SouthAfrica/News/3-held-for-robbing-cyclists-20130326.

Sparks, RF (1982). *Research on Victims of Crime: Accomplishments, Issues, and New Directions*. Rockville, MD, USA: University of California Libraries.

Stutzman, F, & Kramer-Duffield, J (2010). Friends only: Examining a privacy-enhancing behavior in facebook, In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* CHI '10 (pp. 1553–1562). New York, NY, USA: ACM.

The Dutch, Police (2012). Woninginbraak. http://www.politie.nl/onderwerpen/woninginbraak.html.

Tufekci, Z (2008). Can you see me now? audience and disclosure regulation in online social network sites. *Bulletin of Science, Technology and Society*, *28*(1), 20–36.

Waller, I, & Okihiro, NR (1978). *Burglary: The Victim and the Public*. Toronto, Canada: University of Toronto Press.

Wallsten, S (2011). What are we not doing when we're online? Technical report, Technology Policy Institute, Washington DC, USA.

Waters, S, & Ackerman, J (2011). Exploring privacy management on facebook: Motivations and perceived consequences of voluntary disclosure. *Journal of Computer-Mediated Communication*, *17*(1), 101–115.

Weisel, D (2002). Burglary of single-family houses. Technical report, US Department of Justice, Washington DC, USA.

Winchester, S, & Jackson, H (1982). Residential burglary: the limits of prevention. Government document, UK Home Office, London, UK.

doi:10.1186/s40163-014-0008-z
Cite this article as: Stottelaar *et al.*: Online social sports networks as crime facilitators. *Crime Science* 2014 **3**:8.