Cybercrime Data Grand Challenges

i. What is the problem and why is it important? (Material can be harvested from section 1 of the slide-deck)
ii. Why is this difficult to do?
iii .Why is progress possible?
iv. What are the barriers for success?

Overarching Discussion

(i) Bad people do bad things and people get hurt. The cyber world extends and changes the paradigm. Reducing the harm requires evidence-based approaches. Hence we need the evidence: the data!

Jens had good economic model analogy

Socio and technical: both reduce the harm by actors….but also engineer the emerging cyber universe so that harm is reduced.

(ii) Existing data collection is myriad, inconsistent and low quality. The ways cyber permeates lives and spans jurisdictions raises questions of privacy and civil rights norms. We don't have consensus on the desired outcomes. We [want data because we want deeper understanding---and the analytics and theory aren't there. And adversaries are adaptive.

(iii) In analog crime, there has been progress, but it took decades. So we can adapt that here, but we need to be patient. On the private side, enterprises recognize that operational security (as well as all other operations) need to be data-driven. There's receptiveness.

(iv) We need a comprehensive, standard source of data---but that conflicts with the need to go deep, and get ground truth. Standard or diverse?

Grand challenge: how do establish the historical record before the data disappears?

(i)

- The data is there, but it's disappearing.
- We don't understand history.
- How do we insulate collection and preservation of data against the day-to-day pressures of enterprise mission?
- How can we effectively re-contribute seized data?

(ii,iv): Existing work is not longitudinal. Bad quality and methodology. Legal concerns. The data collection itself enables new attacks (iii) Many barriers are policy, not technology; storage is cheap; automation is improving.

Grand challenge: what data needs to be gathered for effective cybercrime statistics?

(i)

- Given the multiplicity and evolving nature of norms, what is a "crime"?
- Discovering and understanding dependencies may require looking at much larger families of cyber-behavior---not just criminal.
- We don't want to measure technical artifacts but "ground truth."
- How do we measure "ground truth" when the underlying technology keeps going through revolutions?

(ii,iv) Uncertain and evolving norms; uncertainty about what constitutes relevant behavior; technology revolutions. (iii) Well, look at the Wayback Machine

Grand challenge: how to develop effective analytics for cybercrime data?

- The state of practice has been emphasizing data, not analytics.
- Predictive analytics would be so useful. How do we get there?
- How do we cope with adaptive adversaries---who will *also* use analytics?
- The data is just the beginning---to solve the problems, we need to understand more. (e.g. what design patterns made the exploited vulnerabilities possible, and why some vulnerabilities at some sites were exploited but not others)

(i,ii,iv) see bullets. (iii) Tremendous low-hanging fruit.

Grand challenge: how do we balance privacy with collection and analysis of this data?

(i,ii,iv)

- How do we accommodate social and cultural norms regarding the data and measured behavior?

- Legal framework needed! Not just crypto.
- How do we enable effective analytics while avoiding the risks of breaches and rogue insiders at the data collection and analytic sites?

(iii) We already collect and analyze sensitive information. (E.g. gov, CMU passwords, Census Bureau)

Grand challenge: how can we make the “prosecution paradigm” (from non-cyber crime) effective in the cyber domain---and what are its limits?

(i,ii,iv)

- How do we accommodate the varying legal norms regarding behavior, data, and evidence?
- How can we reliably solve the *attribution* problem?
- The legal system has long-established standards for human witnesses---how do we develop effective standards for “software on the witness stand”?
- Can the prosecution paradigm scale to the cyber world?
- It takes a HUGE amount of time. People-intensive, not automated. Potential snappy challenge: build the magic machine that does this.

(iii): there is more prosecution than there used to be. There have been advances in bringing analog legal norms into the cyber age (e.g. cellphones, social media)
[DIscussion: distinctions between civil and criminal]

Grand challenge: how to define effective norms of behavior for nation-state actors in cyberspace

- There’s a lot of existing work here; e.g., the “Tallinn Manual” from NATO
- Another dimension: We’ve been thinking about “traditional” hackers. We might be thinking about nation-states. But it’s not a crisp distinction. Contractors.
- What makes a norm “effective”?
- What about rogue states or failed states? Asymmetric actors.

(i) unconstrained competition is a problem; e.g. the “Zero Days” film. (ii,iv) unclear distinctions: what’s an attack and what’s a defense? Espionage vs cyber attack. Absence of consensus.
(iii): Increased nation-state awareness of the risks (and some increased cooperation wrt cybercrime)

Grand challenge: stop the damage from weaponized information

(i)

- Nation-state component
- Not a solely technical problem
- Not necessarily fake.
- Causes real harm
- Doesn’t clearly fit existing norms of criminality
- “Using cyber means for cost-effective social control at scale by an adversary”

(ii, iv) who decides what is “adversarial” and how do they do that? How do we get access to the infrastructure where it happens? Feeds on cognitive biases not overcome by the “obvious” solution of pointing out data is incorrect. (iii) We have made progress on Sybil attacks. We have also gotten good enough at NLP that automation may help scale defenses to match the scale of the attackers. [Discussion: similarity to doxing and online harassment; is “stop” too strong?]

Grand challenge: stop ransomware

- Advanced OS instrumentation might help---but at what privacy cost?
- Better hygiene, and better backups?