Resilience in Cyber-Physical Systems

Weaponizing disinformation to attack critical infrastructure systems

Dr. Jimmy Chih-Hsien Peng

Assistant Professor
Department of Electrical and Computer Engineering

www.penglaboratory.com
Presentation outline

• Critical infrastructure as a target

• Disinformation attacks on power distribution systems

• Disinformation attacks on urban traffic networks


Critical infrastructure as a target

- Critical infrastructure are large **safety-critical systems** whose incapacitation can debilitate national security, the economy, and public health or safety.
  - Power grid, transportation, communications, IT, healthcare, defence, etc.

- These are **cyber-physical systems**, and interact with humans to perform designated functions.


- Critical infrastructure are increasingly becoming a **platform of conflict**.

- Experience diverse and evolving adversarial threats in day-to-day operations.

- Much focus on vulnerability identification and developing protection strategies, mitigation plans.
Facets of critical infrastructure security

- Physical security of assets.
- Ensuring redundancy.
- Designing fundamentally more secure hardware, e.g., hardware-level encryption, no active control ports accessible, MAC-based access control.

- Innate biases, unpredictable.
- Weakest link in real-world examples: e.g., Stuxnet attack on Iran in 2007, Ukrainian power grid attack, 2015 [spear-phishing of operators].
- Organizational changes to inculcate safety culture.

- Firewalls.
- Encryption.
- Antivirus software.
- Regular bug fixes, updates.

- Electricity consumers in power grids, drivers in traffic networks.
- Indirect impact on system when compared to system operators’ actions.
- Possibly be manipulated, at scale, using disinformation.
- Adversary may circumvent existing protection schemes.
The global threat of disinformation

➢ Disinformation refers* to “false information that is purposefully spread to deceive people”.

➢ Disinformation is increasing in scale:
  ▪ Social media platforms promote virality over veracity.
  ▪ Advances in machine learning, and social bots aid in spreading it at a large scale.

➢ Disinformation is becoming more sophisticated:
  ▪ Narrowly tailored to fit the target audience.
  ▪ Varied manifestations such as counterfactual social media posts, manipulated news stories, and deep fake videos.

➢ Can be weaponized to manipulate long-term decisions of a society, such as elections’ outcomes, and wars.
  ▪ E.g., prevalence during 2016 US/Brexit elections, ISIS capture of Syria.

*Source: Lazer et al., "The science of fake news", Science, March 2018
Motivation for this research

➢ The possibility that a malicious actor could use disinformation in a targeted attack to influence social behaviour within a **limited time span** has not been considered to date.

➢ Unstudied vulnerability: decisions of individuals in complex cyber-physical systems could be manipulated to the overall detriment of the system.

   ▪ Engineers assumes some degree of **rationality and predictability** of user behaviour, **at least collectively at the system-level**.
   ▪ Resilience may not be guaranteed when individuals behave irrationally and unpredictably.
   ▪ Individuals may not even be aware of the impact of their actions at the system level!

Can disinformation be used to compromise the security of critical infrastructure?
Bringing social behaviour into the loop

Traditional analysis

Cyber-physical infrastructure

Operation and control

Information/Feedback

Society-in-the-loop analysis of resilience

Monitoring

Social behaviour
Questions to answer

➢ Focus on two critical infrastructure systems:

1. Power distribution system.
2. Urban road traffic.

Their performance highly dependent on social behaviour!

1. What are the mechanisms of a behaviour manipulation attack targeting the power grid and urban traffic networks?
   • Alternatively, what disinformation designs could manipulate the masses in each system to behave to the detriment of the system?

2. Since the impact of a disinformation attack depends on the number of people reacting to the disinformation, can we estimate what fraction of recipients would do so in reality (i.e., the follow-through rate)?

3. Given a disinformation follow-through rate, can simulations be developed to quantify the impact of the attack?

Overall, can disinformation campaigns bring down critical infrastructure systems such as the power grid and urban traffic networks?
Disinformation attacks on the 

POWER DISTRIBUTION SYSTEM


An adversary could manipulate consumer behaviour and alter the demand patterns:
Disinformation design and delivery

➢ The adversary could:
  ▪ Hijack DR applications/communication channels to manipulate the content.
  ▪ Use *spoofing attacks* to spread disinformation:
    ▪ No hack of utility infrastructure, hard to detect.
    ▪ Consumers have no reason to suspect DR messages to be fake! ~As long as fidelity to legitimate messages is preserved.

➢ Effective disinformation will be *tailored to fit the scheme of legitimate communications* from the utility, which are usually published online*.

➢ Only *residential consumers* are considered in this study.
  ▪ Human decision-making in C&I consumers low.
  ▪ Residential loads less critical.

Modelling consumer response-BAFT model

Belief status: \( \theta_b = \begin{cases} 1, & \text{if resident believes the notification to be authentic} \\ 0, & \text{otherwise} \end{cases} \)

Propensity to Accept events: \( \theta_a \in [0, 1] \)

Propensity to Follow-Through on the event task: \( \theta_{ft} \in [0, 1] \)

Overall follow-through rate: \( \theta = \theta_b \theta_a \theta_{ft} \)

Translating to appliance-use probabilities

\[ P_{t,i}^* = P_{t,i} (1 - \theta \theta_{util}), \forall t \in [t_{start}, t_{end}], A_i \in A_{dr} \]

\[ P_{\text{deferred},i} = \sum_{t=t_{start}}^{t_{end}} (P_{t,i} - P_{t,i}^*). \]

IEEE standard 123-node test feeder case

- 2094 residential consumers

Varying DR penetration level

Varying Acceptance & Follow-through rates
Attack - Fake DR notifications

- The adversary schedules fake DR events so as to increase the overall system peak demand.
- If event period is just before the peak period, the overshoot after the event increases the system peak demand.

"Please do not consume more than 3kWh tomorrow between 5-7PM. Doing so, earn a FREE lottery ticket! TIP: For best results, just schedule your washing machines, dryers and dishwashers to work between 7-9PM."

Peak demand period: 8-10PM
**Attack - Fake DR notifications**

- Variation of maximum aggregate system demand and minimum voltage:
  - Minimum voltage decreases, and system peak demand increases.
  - As more participants believe in the adversary’s message, the more is the effect on the network.
Attack- Fake maintenance shutdown alerts

➢ Fake maintenance alerts may be spread by the adversary.

“Maintenance activities could affect electricity supply between 9PM and 12AM; try to use any appliances before the maintenance period.”
Attack - Intercepting legitimate DR messages

- Legitimate DR messages intercepted by attacker. Alternatively, fake message declaring event cancellation are sent out.

- Adversary may avert utility suspicions by sending false acceptance notifications to the utility.

- Results in **unexpectedly low reserves** during the peak demand period.
Strategic response of the utility

➢ **Stackelberg non-cooperative game** to model utility countermeasures to nullify the attack impact.
  ▪ Adversary: leader/attacker,
  ▪ Power Utility: follower/defender.

➢ If attack is detected before consumers react, the utility can broadcast counter-messages.

➢ **Payoffs:**
  ▪ If no attack happens, both players receive zero payoff.
  ▪ If defender does not detect the attack, the defender receives a negative payoff and attacker a positive payoff.
  ▪ If attack is detected, the payoff to defender depends on the time of detection, the consumer response to its counter-message, as well as the potential impact of the attack if it were undetected.

➢ Solve for the Stackelberg equilibrium, which **maximizes the attacker’s expected payoff**.
The Stackelberg equilibrium

- Immediate detection with the threat of effective counter measures can deter even the most sophisticated attacks! ~Effective detection mechanism required.

- Late detection means that attacker will attack near the peak demand period (8PM, strategy 21), unless the cost of attacking is prohibitively high. ~Fast detection essential.
Estimating disinformation follow-through rates

Participants were shown 4 types of messages:

- **Message 1**: Today only! Click here to enjoy a discount of 50% on your electricity rate from 8PM to 10PM. Spread the word!
- **Message 2**: Today only! Enjoy a discount of 50% off your electricity rate from 8PM to 10PM. Spread the word!
- **Message 3**: Today only! Click here to enjoy a discount of 50% on your electricity rate from 8PM to 10PM. Spread the word!
- **Message 4**: Today only! Enjoy a discount of 50% off your electricity rate from 8PM to 10PM. Spread the word!

**Models of propagation**
- Linear Threshold
- Independent Cascade

**Survey (n = 5,124)**
Influence propagation simulations

Survey results | Probability mapping function | Network model (1 million nodes) | Influence propagation model

Follow-through propensity
Forwarding propensity

**Probability mapping function**

- **Linear**
- **Squared**
- **Cubic**

**Network model**

- Prefential attachment networks (Barabási-Albert model)
- Random graphs (Erdős-Rényi model)
- Small world networks (Watts-Strogatz model)
- Scale-free networks (Newman Configuration model)

**Influence propagation model**

- Linear Threshold
- Independent Cascade
- Final follow-through rate
We only consider 1 round of propagation.

- A previous study* analysed 1 billion diffusion events on Twitter.
- Most diffusion events terminate at the root itself, or after 1 round.

Final follow-through rates range from 3.2%—26.8%.

The removal of a link always increases the follow-through rate.

- 3.4%—9.8% increase in follow-through after 1 round of propagation.
- Difference between phishing attacks and our scenario—no link is required!

Case study- Greater London distribution network

Transmission level substations feeding Greater London (data from National Grid)

Geographical and building data (from OpenStreetMap)

Distribution network topology of Greater London

- 9 “district networks” or subnetworks.
- 398,266 residential consumers.
Impact of disinformation “follow-through” on residential load profiles

Disinformation notification received by consumers

The peak demand increases due to the attack, potentially leading to line overloads.
Impact of manipulated consumer behaviour

➢ Simplified power system analysis by focusing on real power flows and discarding voltage and reactive power.

▪ Previous studies* of real US and UK grids show power line capacity limit is the first bottleneck, and not the voltage limits.

➢ The peak power capacity of the distribution feeders is limited, i.e., circuit breakers trip the system if power/current limits are reached.

\[
\frac{(P_{\text{peak, attack}} - P_{\text{peak, normal}})}{(P_{\text{peak, normal}})} > \text{Threshold}
\]

➢ An attack that results in a demand exceeding of the above threshold leads to the tripping of the corresponding feeder, and all lines downstream of it.

➢ Capacity constraint further strained by the inclusion of new high-power loads such as Electric Vehicles.


Impact of the attack on the power grid

➢ We simulate the attack on the Greater London power grid, with 10% threshold.

➢ Blackouts are dispersed across the network rather than concentrated:
Impact of grid upgrades

\[
\frac{(P_{\text{peak, attack}} - P_{\text{peak, normal}})}{(P_{\text{peak, normal}})} > \text{Threshold}
\]

➢ This term includes both residential appliance + EV demand under “no attack” conditions.
➢ This means we “upgrade” the grid to support EV adoption for each simulation.

If upgrades are delayed, impact further worsens:
➢ Delaying upgrades is equivalent to reducing the overloading capacity of the feeders.
Impact of the attack on the power grid

➢ Consider 30% of population targeted initially. Heatmap could be used as a lookup table to ascertain grid impact.

➢ Final follow-through rates range from 9.4%—26.8%.

➢ Depending on the EV penetration level, the size of the population affected by the blackout will be 5.6%—100%.
Implications

➢ By using strategic and well-tailored disinformation, an adversary can attack the power system without directly targeting its software or hardware, or operators.

➢ Survey showed people not only willing to follow-through, but also forward notifications, thereby amplifying the attack.

➢ **Heavily-loaded grids** are particularly vulnerable.
  - Impact is likely to worsen as the available flexible demand increases.
  - As demand is seasonal, grid is more vulnerable when the demand is highest.

➢ **Timely grid upgrades** of grid infrastructure capacity are necessary.
  - Delays in upgrades increase congestion during normal operation, and the impact of a potential attack.

➢ Attack **detection and mitigation** strategies must be developed.
  - There is a window for the utility to detect and send out countermeasures before consumers respond.
  - Need to increase awareness and immunize the public to disinformation.
Disinformation attacks on URBAN ROAD TRAFFIC

Manipulating drivers’ behaviours and traffic flows

- Resulted in traffic spilling over and causing gridlocks on local streets near the area.
- Could such an incident be manufactured, but without actually blocking the lanes?

What if people were made to believe there was an accident/heavy congestion in a location when there was none in reality? What if people were attracted by a “honeypot” to one particular location?

Spoofing Google’s algorithms with a wagon full of smart phones with Google Maps turned on

Image source: NBC news

Brand name: NUS National University of Singapore

© Jimmy Chih-Hsien Peng
Manipulating drivers’ behaviours and traffic flows

➢ Two broad types of attack mechanisms are studied here:

Divergence attacks redirect traffic away from a location(s) in the city

Convergence attacks redirect traffic to a location in the city
Divergence attacks

➢ Two mechanisms are considered:
  ▪ Fake accident/congestion notifications.
  ▪ Physically placing fake road closure signs at the side of certain roads.
Convergence attacks

- Inspired by the real incident in Dubai.
- Fake alerts advertising a limited-time, massive sale at a popular retail store.

Direction of one-way street
Location targeted in convergence attack
Possible path taken by drivers following-through
Path taken by drivers not following-through
How many people would believe such disinformation?

**DIVERGENCE ATTACK**

- 89% of participants report *follow-through* propensity greater than 5/10 for traffic alerts, 97% for road signs.
  - Majority of drivers’ behaviour would be altered by a divergence attack.

- 55% of participants report *forwarding* propensity greater than 5/10.
  - People more likely to follow-through rather than forward.
  - Overall disinformation reach expected to increase due to forwarding.

**CONVERGENCE ATTACK**

- Responses lower than that of divergence attack, yet worrisome.
  - 50% of participants report *follow-through* propensity greater than 5/10.
  - 47% of participants report *forwarding* propensity greater than 5/10.

- Yet, smaller follow-through rates can cause major disruptions in this attack as all vehicles head to the same location.
Chicago case study

Street data from OpenStreetMap

Traffic simulations in Chicago

One-way streets
Two-way streets

Historical traffic data for Chicago

1. Daily average number of vehicles passing by different locations in the city.
2. Distribution of the intensity of the traffic throughout the day.

✓ Average ride time in our simulations: 22.59 min
   Statistics reported in the American Driving Survey: 23.18 min

✓ Generated rides match the locational data.
✓ Generated rides match the daily traffic intensity distribution.
Attack impact on streets

➢ Consider a divergence attack on 10 targets.
  ▪ Each direction of traffic in a street is a potential target.

➢ Targets selected by the greedy heuristic are spread across the city, and not in one neighbourhood.

➢ Traffic decreases in some streets, and increases in others. ~Total number of rides remain the same.

➢ Impact on traffic propagates from the targets.
Attack impact on streets

- **Convergence attack** on “Target” retail store in downtown Chicago.

- Rides passing within 1 km radius of this store may follow-through and change their routes to pass by the store.

- 1,000 extra rides from random starting nodes to the store are also added.

- Disruption centred around the targeted location, and impact reduces as distance to the target decreases.

Follow-through rate = 10%
Attack impact on drivers

DIVERGENCE ATTACKS:

- Divergence attack using **greedy heuristic** on 10 targets. For comparison, we also depict results from an attack where targets are randomly selected.

- Ride time increases for some drivers, decreases for others.
  - **Braess’ paradox**: removal of some streets reduces ride times!

- Distribution skewed to left for greedy heuristic, to the right for random attack: e.g., for 50% follow-through:
  - Greedy: 9.28% delay, 7.33% speed up.
  - Random: 5.49% delay, 5.57% speed up.

  *Attack strategy is important.*

- Higher follow-through rate increases variance of the time-delay distributions.
  - More rides suffer larger delays.
**Attack impact on drivers**

**CONVERGENCE ATTACKS:**

- Ride time becomes longer for some, and shorter for others.
- Do only rides that come close to the target slow down?
  - Insets show rides that do not pass within 1km radius of the target.
  - Distribution in the inset not skewed regardless of follow-through rate.
  - Vast majority of these rides experience almost no delay (note Y axis is in log scale).
Implications

➢ Two types of disinformation attacks on drivers. Disinformation mechanisms include fake traffic alerts, road signs, and store discount notifications.

➢ Surveys show potentially high follow-through rates possible, resulting in significant disruptions.

➢ Though achieving maximum impact in a divergence attack is computationally intractable, even a simple heuristic results in far-reaching disruption. Further, its impact can be focused on a critical neighbourhood of the city.

➢ It is important to detect and counter such disinformation.
  - Can crowdsource the verification of veracity of traffic incidents.
  - Waze offers this feature, but serves only a small consumer base (11% of all users within the US).
  - Policy implication: must extend this functionality to all navigation apps.
Summary

➢ Disinformation campaigns not only influence wars and politics, but also the security of critical infrastructure systems by manipulating the behaviour of the people at large.

➢ Vulnerabilities in critical infrastructure systems can arise not only from hardware, software, and human operators, but also from the large-scale manipulation of individuals’ behaviour. That is, information security/cybersecurity risks can cascade into other forms of risk.

➢ Impact of attack may be magnified through propagation in social networks.

➢ Tackling these requires explicit modelling of consumer behaviour.

➢ Detection and mitigation of such disinformation attacks essential.
  ▪ Immunizing the public to disinformation.
  ▪ Timely counter-messaging.
  ▪ Timely upgrades of power grid infrastructure (power systems case).
  ▪ Allowing crowdsourced reporting of false traffic information on navigation apps (traffic networks case).
Science based risk management and community resilience

Anshu Sharma
04 August 2021
3 elements of resilience

- Anticipate disasters through better risk understanding
- Absorb impacts during the shock
- Adapt through learning and improvement after the experience
Anticipate
Measuring locally

- Automated Weather Stations – accurate, 24x7, reliable, easy to monitor
- Useful for localising forecasts with support from met agencies

BUT
- Expensive, invisible and sterile
Understanding locally

- Climate school
- What will happen if it rains a lot?
  BUT
- Intense on human engagement, needing local commitment
Engaging locally

- So what can we do?
- Conversations on community radio, engineered by local youth groups

BUT

- Need institutionalisation to sustain
Why low tech is better

- Sunny Weather Labs
- Look, feel, touch, operate
- Create local weather-men and women
Managing Complex Disasters, Sikkim

- Ministry of Environment, Forests and Climate Change
- Sikkim State Disaster Management Authority
- Indian Institute of Public Administration
- SEEDS
# Weather Watch

**SCHOOL NAME:**

<table>
<thead>
<tr>
<th>Today’s Weather</th>
<th>DATE:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. Temp.</td>
<td>°C</td>
</tr>
<tr>
<td>Min. Temp.</td>
<td>°C</td>
</tr>
<tr>
<td>Rainfall</td>
<td>mm</td>
</tr>
<tr>
<td>Humidity</td>
<td>%</td>
</tr>
<tr>
<td>Wind Direction</td>
<td></td>
</tr>
<tr>
<td>Wind Speed</td>
<td>km/hr</td>
</tr>
<tr>
<td>Atmospheric Pressure</td>
<td>mm Hg</td>
</tr>
</tbody>
</table>

**Monthly Average Weather**

<table>
<thead>
<tr>
<th>MONTH:</th>
</tr>
</thead>
<tbody>
<tr>
<td>YEAR:</td>
</tr>
<tr>
<td>Max. Temp.</td>
</tr>
<tr>
<td>Min. Temp.</td>
</tr>
<tr>
<td>Rainfall</td>
</tr>
<tr>
<td>Humidity</td>
</tr>
<tr>
<td>Wind Direction</td>
</tr>
<tr>
<td>Wind Speed</td>
</tr>
<tr>
<td>Atmospheric Pressure</td>
</tr>
</tbody>
</table>
Local weather advisory
Mapping flies, mosquitoes, garbage
Community Risk Register

Risk Matrix to be filled based on Likelihood and impact scoring
Government ‘ownership’

- Secretary and Relief Commissioner, Sikkim, at Maghe Mela
Tatkal Apada Report

Give nearest disasters information

Get relief help as per disaster details provided
POOR RESOLUTION OF EXISTING VULNERABILITY MAPPING

Understanding Risk is Challenging as Existing Hazard Warnings are Not issued for Particular Locations and Vulnerability Maps Fail to Consider Specific Area and Population Based Vulnerabilities

SUNNY LIVES: AI FOR HUMANITARIAN ACTION

AI-Enabled Model that Analyses Large Volumes of Localised Data Sets to Provide Real-Time Hazard Warnings and Risk Advisories for Marginalized Communities
1. Organise better to manage the disaster (governance)
2. Educate communities better
3. Understand underlying risks
4. Prepare
5. Absorb the shock better – saving lives, assets and productivity
6. Apply learnings from experience to adapt