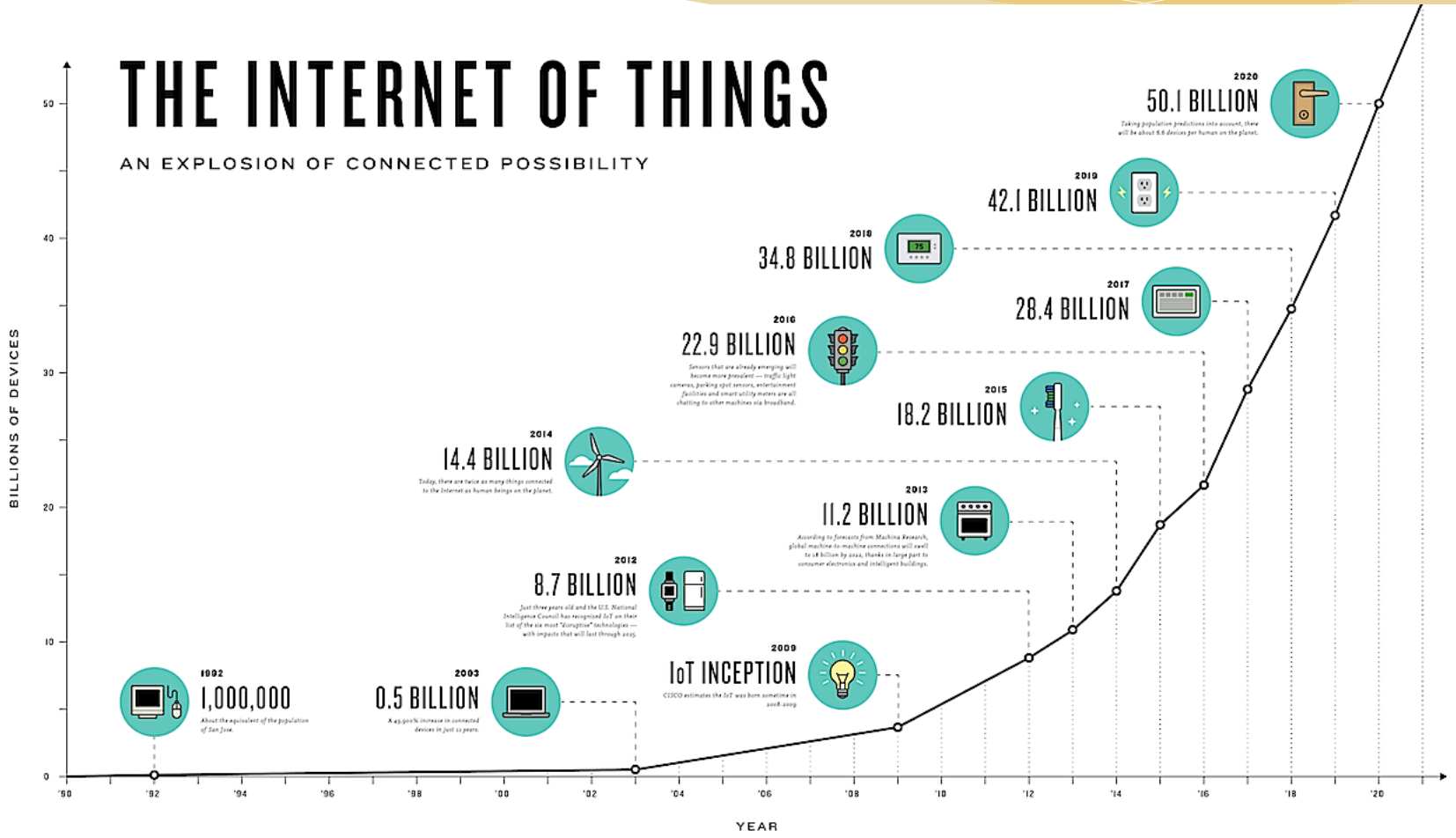
FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

# Neural Network-based Graph Embedding for Cross-Platform Binary Code Similarity Detection

Chang Liu
UC, Berkeley
FORCES, 2017

Berkeley UNIVERSITY OF CALIFORNIA
Massachusetts Institute of Technology
NSF
MICHIGAN
VANDERBILT UNIVERSITY






FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

# Firmwares of IoT devices

* Binary code is in various ISA
  * X86, MIPS, ARM, etc.

* Problem: the current practice is to employ common open sourced code when building firmware; thus one vulnerability identified in the source code may affect millions of firmwares
  * E.g., Heartbleed in OpenSSL

# Vulnerability Detection

* Goal: given a (newly detected) vulnerability, quickly identify whether it affects an existing firmware.

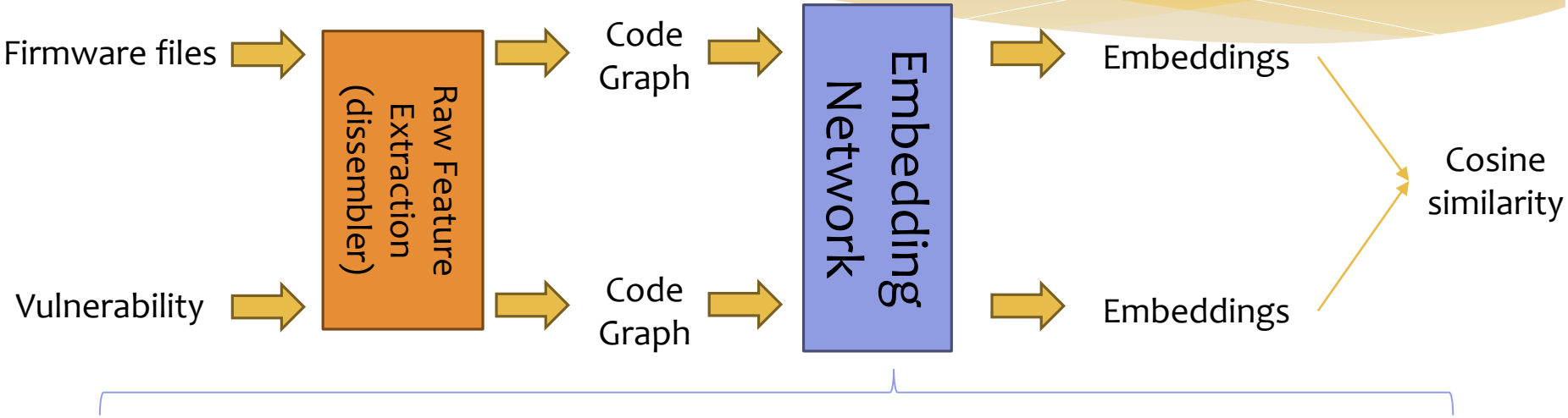
* **Xiaojun Xu**, **Chang Liu**, Qian Feng, Heng Yin, Le Song, **Dawn Song**, *Neural Network-based Graph Embedding for Cross-Platform Binary Code Similarity Detection*, to appear in ACM CCS 2017

# Deep learning for vulnerability detection

- Problem definition:
  - Given two programs:
    - One is the vulnerability (e.g., compiled from pre-patched vulnerable source code)
    - The other is a binary code
  - Detect whether the two programs are similar

- Challenges:
  - The ISAs of the two programs may not be the same
  - Compiler and compilation options (e.g., optimization levels) may not be the same
  - Collision-resilient hash function-based detection is far from accurate

# Overall workflow




Previous approaches

- Manually designed graph-matching-based algorithms
- Slow
- Effectiveness is limited by graph-matching
- Feng, et al. Scalable Graph-based Bug Search for Firmware Images. *CCS* 2016.

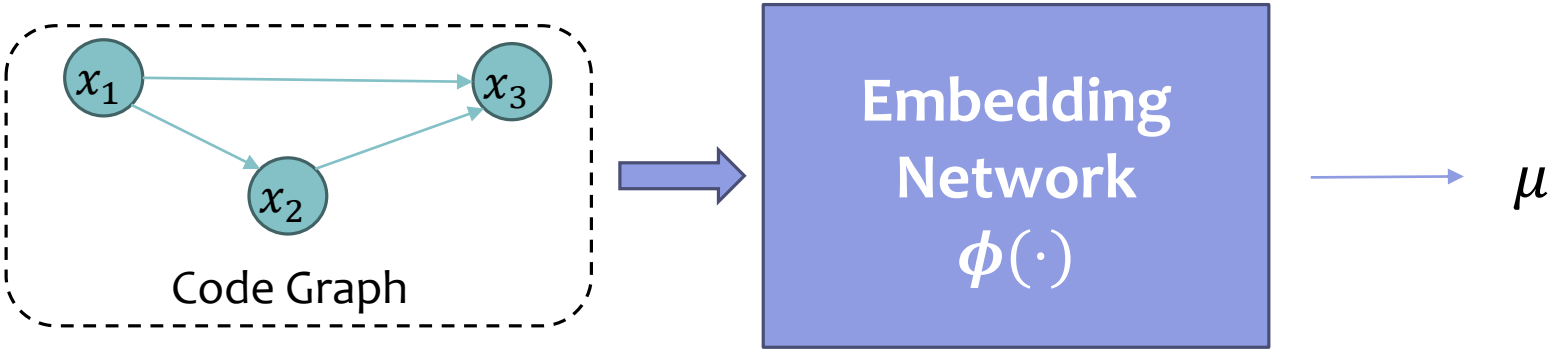
Our approaches:

- Deep graph embedding network
- Design a neural network to extract the features automatically
- Combine Struct2vec and Siamese network



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

# Our approach: struct2vec

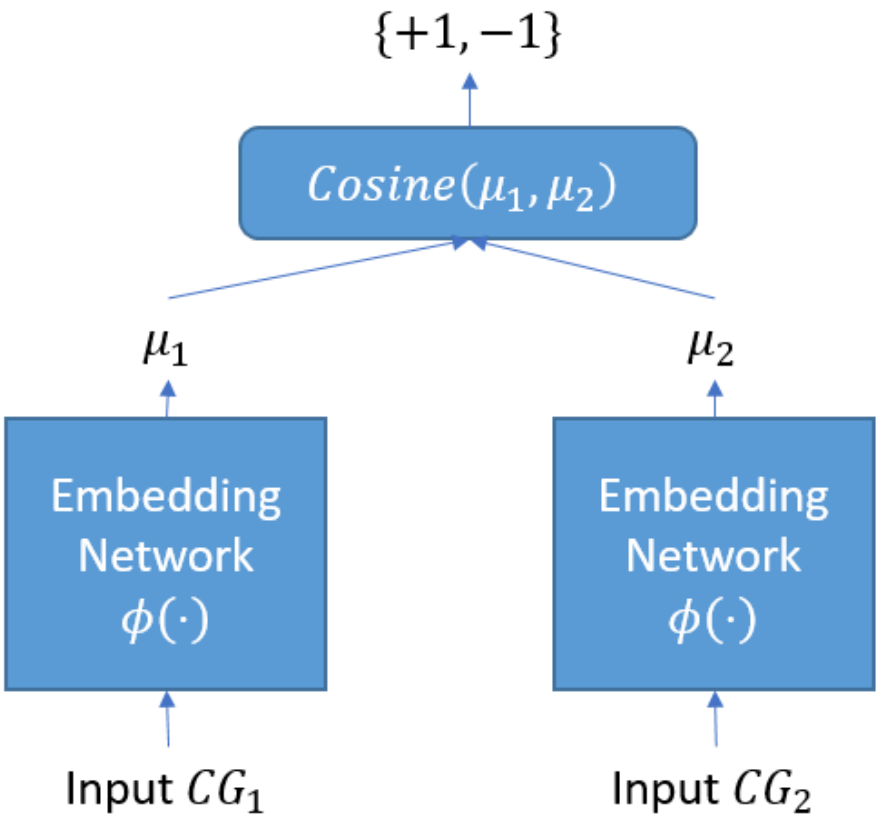



Dai, et al. Discriminative Embeddings of Latent Variable Models for Structured Data. ICML 2016.



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

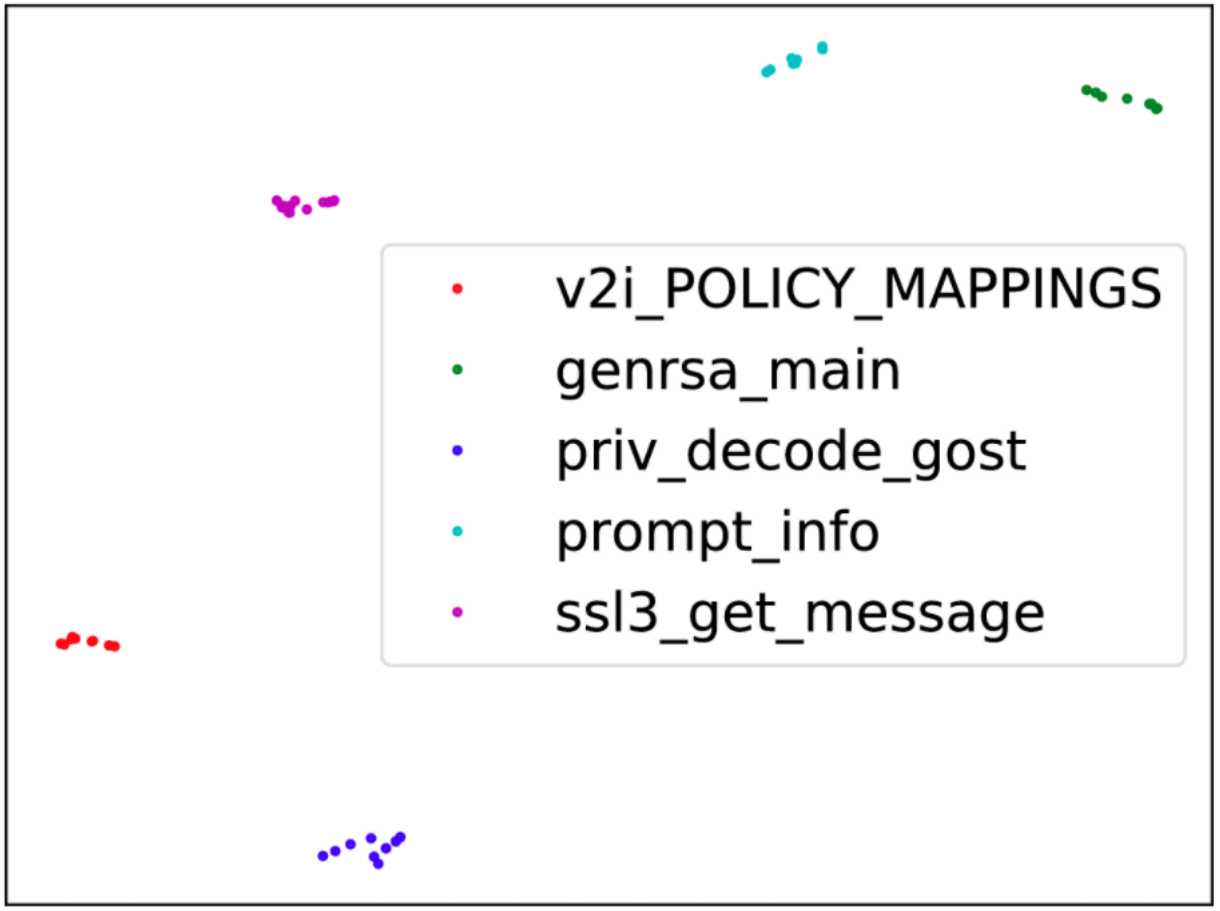
# Training: Siamese






FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

# Visualizing the embeddings






FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

# Serving time (per function processing time)

Previous work: a few secs to a few mins

**Now: a few milliseconds**

**2500 × to 16000 × faster!**

# Training time

Previous work: > 1 week

**Now: < 30 mins**

# Identified Vulnerabilities

| Function Name | Vendor | Firmware | Binary File | Similarity |
|---|---|---|---|---|
| ssl3_get_new_session_ticket | D-Link | DAP-1562_FIRMWARE_1.00 | wpa_supplicant.acfgs | 0.962374508 |
| port_check_v6 | D-Link | DES-1210-28_REVB_FIRMWARE_3.12.015 | in.ftpd.acfgs | 0.955408692 |
| sub_42EE7C | TP-Link | TD-W8970B_V1_140624 | racoon.acfgs | 0.954742193 |
| sub_42EE7C | TP-Link | TD-W8970_V1_130828 | racoon.acfgs | 0.954742193 |
| prsa_parse_file | TP-Link | Archer_D5_V1_140804 | racoon.acfgs | 0.949814439 |
| sub_432B8C | TP-Link | TD-W8970B_V1_140624 | racoon.acfgs | 0.949583828 |
| sub_432B8C | TP-Link | TD-W8970_V1_130828 | racoon.acfgs | 0.949583828 |
| ssl3_get_new_session_ticket | DD-wrt | dd-wrt.v24-23838_NEWD-2_K3.x_mega-WNR3500v2_VC | openvpn.acfgs | 0.94668287 |
| ucSetUsbipServer | TP-Link | WDR4900_V2_130115 | httpd.acfgs | 0.946312308 |
| ssl3_get_new_session_ticket | Netgear | tomato-Cisco-M10v2-NVRAM32K-1.28.RT-N5x-MIPSR2-110-PL-Mini | libssl.so.1.0.0.acfgs | 0.945933044 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-K26-1.28.RT-MIPSR1-109-Mini | libssl.so.1.0.0.acfgs | 0.945933044 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-K26USB-1.28.RT-N5x-MIPSR2-110-VPN | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-E4200USB-NVRAM60K-1.28.RT-MIPSR2-110-PL-BT | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-E3000USB-NVRAM60K-1.28.RT-MIPSR2-110-BT-VPN | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-K26USB-1.28.RT-MIPSR1-109-AIO | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-Netgear-3500Lv2-K26USB-1.28.RT-N5x--109-AIO | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-E4200USB-NVRAM60K-1.28.RT-MIPSR2-109-AIO | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-E1550USB-NVRAM60K-1.28.RT-N5x-MIPSR2-110-Nocat-VPN | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-K26USB-1.28.RT-N5x-MIPSR2-115-PL-L600N | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-E1550USB-NVRAM60K-1.28.RT-N5x-MIPSR2-110-BT-VPN | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-E3000USB-NVRAM60K-1.28.RT-MIPSR2-108-PL-VPN | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-E1550USB-NVRAM60K-1.28.RT-N5x-MIPSR2-110-Mega-VPN | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-E1200v2-NVRAM64K-1.28.RT-N5x-MIPSR2-108-PL-Max | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-K26USB-1.28.RT-MIPSR1-109-Mega-VPN | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-E3000USB-NVRAM60K-1.28.RT-MIPSR2-109-Big-VPN | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-E4200USB-NVRAM60K-1.28.RT-MIPSR2-108-PL-Nocat-VPN | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-Netgear-3500Lv2-K26USB-1.28.RT-N5x--110-ND-AIO | libssl.so.1.0.0.acfgs | 0.945932984 |
| ssl3_get_new_session_ticket | Tomato_by_Shibby | tomato-E4200USB-NVRAM60K-1.28.RT-MIPSR2-109-Nocat-VPN | libssl.so.1.0.0.acfgs | 0.945932984 |

Among top 50: **42** out of **50** are confirmed vulnerabilities

Previous work: **10**/**50**



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

# Takeaways

*Message 1*. Deep learning approaches can be not only **more effective**, but also **more efficient** in learning embedding representations for binary **programs**.

*Message 2*. Program analysis can be a **novel application domain** of deep learning techniques toward a more **secure cyber-physical world**.