URL Semantic Analysis for Phishing Detection

Presentation ResCom 2014
May 13, 2014

Samuel Marchal
Ph.D. student
samuel.marchal@uni.lu

SnT – Interdisciplinary Centre for Security, Reliability and Trust - Luxembourg
TELECOM Nancy – University of Lorraine - France

What is Phishing ?

securityandtrust.lu

- Use of social engineering and technical subterfuges to steal consumers' data:
  - Identity information
  - Web-sites credentials
  - Credit card information
  - Etc.

- Cause billions of dollars of loss every year

- SoA detction techniques
  - Reactive blacklisting (*e.g.* PhishTank)
  - Web-site graphical analysis
  - Passive DNS analysis
  - URL lexical analysis

Observation (phishing)

Source: APWG

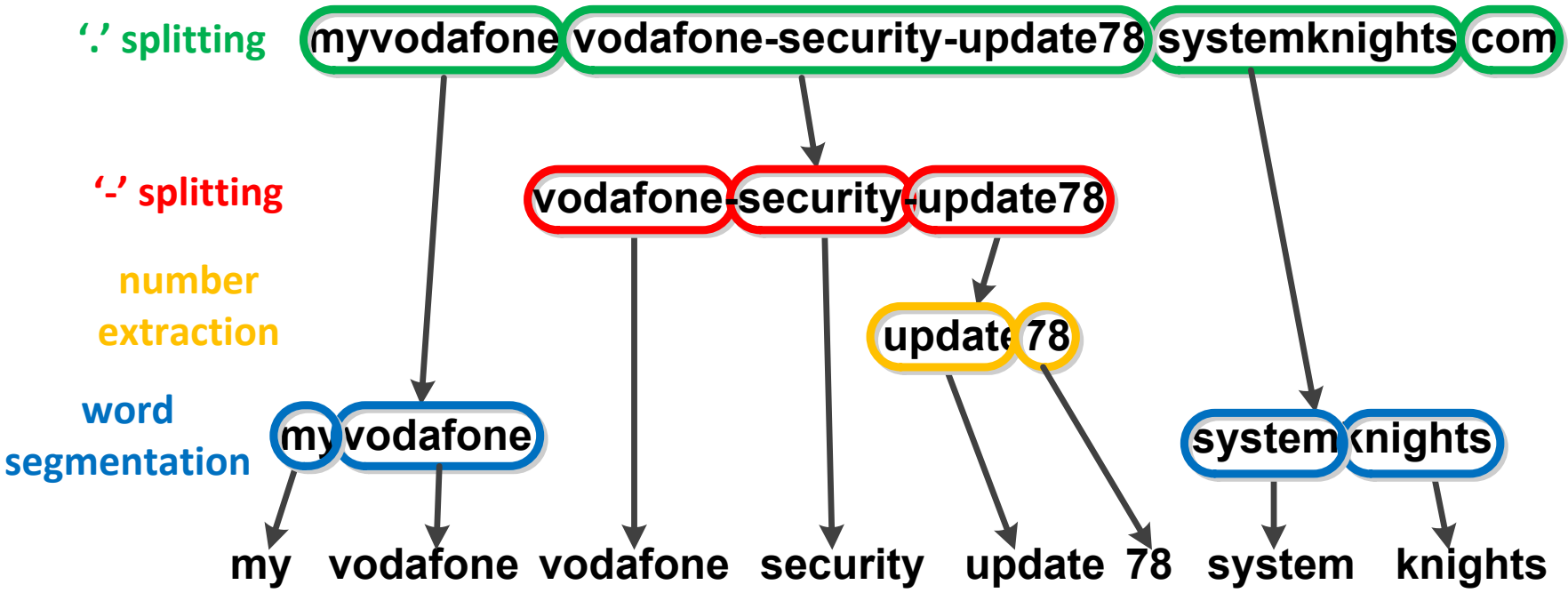
securityandtrust.lu

- Mainly relies on social engineering (easier)
- Use obfuscated URLs with similar patterns:
  - Targeted brands
  - Attractive keywords (secure, login, etc.)
- URLs are meaningful

Idea

→ **Perform a semantic analysis of URLs To detect phishing**

URL pre-processing



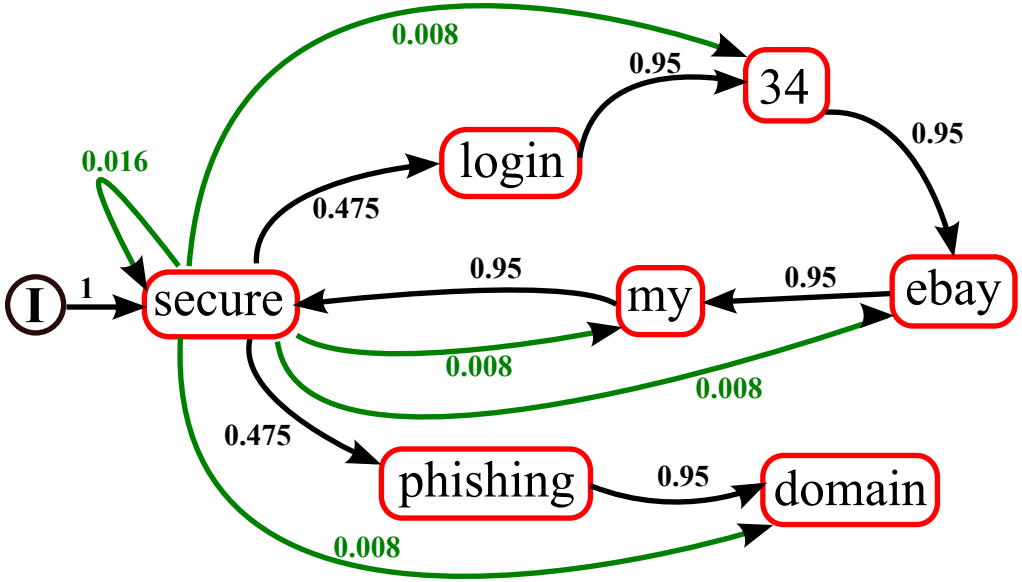
Set of words composing the URL

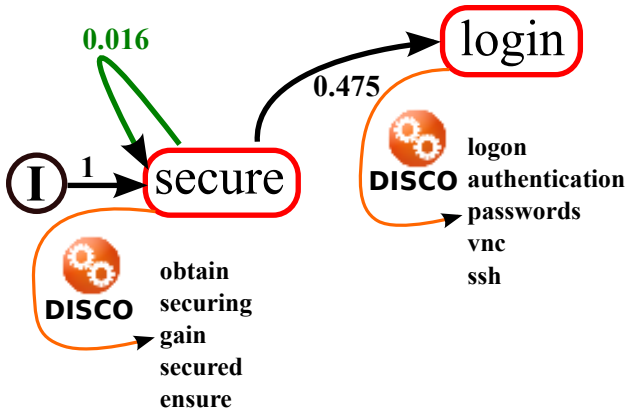
Applications

securityandtrust.lu

1. Phishing domains prediction (*i.e.* automaticaly generate domains likely to be registered by phishers)

- Study the composition of several known phishing domains
- Build a Markov chain representing their compostion pattern
- Extend the genration model with semantic tools





Applications

securityandtrust.lu

2. Identify URL set maliciousness

- Group set of URL according to common features (*e.g.* pointing to a common @IP)
- Semanticaly compare with labeled sets of URL (legitimate / phishing)
- New metrics proposed to quantify semantic similarity between two sets of words

	leg-5	leg-4	leg-3	leg-2	leg-1	mal-5	mal-4	mal-3	mal-2
mal-1	0.776	0.795	0.793	0.789	0.785	0.955	0.962	0.965	0.975
mal-2	0.782	0.800	0.798	0.797	0.797	0.965	0.968	0.973	
mal-3	0.772	0.796	0.793	0.788	0.784	0.951	0.962		
mal-4	0.783	0.804	0.804	0.800	0.796	0.953			
mal-5	0.769	0.785	0.784	0.782	0.772				
leg-1	0.946	0.948	0.952	0.938					
leg-2	0.915	0.924	0.922						
leg-3	0.936	0.934							
leg-4	0.935								

URL Semantic Analysis for Phishing Detection

Presentation ResCom 2014
May 13, 2014

Samuel Marchal
Ph.D. student
samuel.marchal@uni.lu

SnT – Interdisciplinary Centre for Security, Reliability and Trust - Luxembourg
TELECOM Nancy – University of Lorraine – France

Thank you