Fifty shades of grey hat: A socio-psychological analysis of conversations on hacking forums

BU Bournemouth University

Dr John McAlaney, Emily Kimpton, Dr Helen Thackeray
Email: jmcalaney@bournemouth.ac.uk

Background

- There is a major recruitment crisis within the cybersecurity industry [1]; yet at the same time there is an active community of people with an interest in hacking and cybersecurity who interact with each other through online forums and offline conventions
- Conversations held within these settings cover the full range of hacking activities and attitudes from overtly criminal (black hat) to purely ethical (white hat) activities, and all of the variations in-between (grey hat)
- Online forums may be the first contact that young people with an interest in computing have with the hacking world, as such it is important to better understand the beliefs, attitudes and group dynamics that they experience and the influence that this has
- Through engaging with these communities we can encourage individuals to pursue the many legitimate careers in cybersecurity where they can follow their passions, and in doing help address the cybersecurity jobs gap

Method

- Data was collected from four online forums and three subreddits (from the Reddit website) related to hacking, identified in part through the engagement the authors have undertaken with hacking communities [2], and following web forum text selection processes recommended by Holtz et al [3]
- Discussion threads were selected based on if they included substantive discussion on the topic of hacking
- Approximately 120,000 words of text were collected for the purposes of analysis
- The first phase of analysis made use of Linguistic Inquiry Word Count (LIWC) software [4], which has been used in previous research to determine the linguistic categories and psychological processes evident in a selection of text
- The second phase of analysis was a thematic analysis of a sample of the text collected from each forum or subreddit

Linguistic Inquiry Word Count (LIWC) analysis

- LIWC software was used to analyse each forum/ subreddit on four measure of analytical thinking, clout, authenticity and emotional tone
  - Analytical thinking reflects the degree to which individuals demonstrate formal, logical and hierarchical thinking patterns
  - Clout is a measure of the relative social status, confidence and leadership that individuals display through their writing
  - Authenticity measures the degree to which people are using language that indicates they are being more personal, honest and vulnerable
  - Emotional tone indicates the degree to which positive and negative emotions are evident in the text, with a score below 50 suggesting a more negative emotional tone

	Word count	Analytic	Clout	Authentic	Tone
Forum 1	8595	33.1	55.59	43.84	51.92
Forum 2	11112	41.73	62.34	52.85	62.96
Forum 3	24986	47.98	52.9	44.7	45.7
Forum 4	19014	54.4	59.89	29.97	44.03
Subreddit 1	14116	54.78	60.41	47.9	55.84
Subreddit 2	25104	53.18	56.78	31.41	58.78
Subreddit 3	15687	51.98	64.98	31.14	47.97

Note the closer the score is to 100 the greater the presence of that variable in the text

Quotes

'As far as self-taught vs otherwise, the difference in this field is passion. People who are self-taught are that way because they have passion for it, and that leads to them becoming smart/successful'

'If you didn't know when unhashing process stops some passwords don't get recovered, so it wouldn't match every right password! YOU ARE A SUCH A NOOB!' [sic]

'Can anyone confirm is this is too risky now? I think I might take a break from it now'

'The information could be biased or plain incorrect, of course not bashing everybody who teaches hacking but simply put a very high percentage of people here don't know a dime about hacking'

'I'm a lot more antagonistic towards script kiddies than I should be, but your eagerness and drive is what will make you guys the next cybersecurity professionals to take over the field'

Thematic analysis

An initial, exploratory thematic analysis procedure was used, following Braun & Clarke (2006). The following interconnected themes were identified:
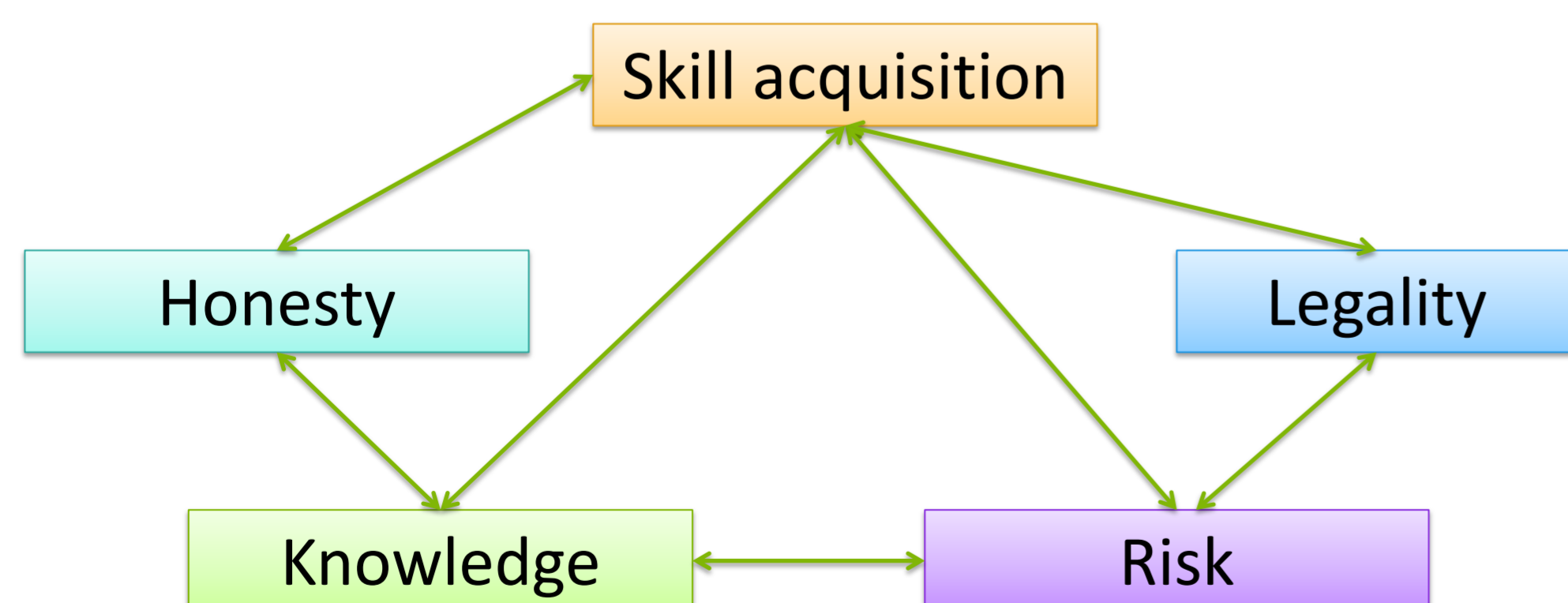
Discussion

As demonstrated by the LIWC analysis there variation evident in the language style and psychological processes used within different sites. Overall the online discussions were characterised by high levels of clout, indicating that users exhibit confidence and preference towards social status and leadership. This was consistent was the themes of knowledge and skill acquisition, which suggested that the intellectual challenges of hacking are a powerful motivator for individuals to become involved in this activity. This is consistent with proposed typologies of hacker identity [5]. Throughout the analysis there was a strong awareness of the potential illegality and risks of hacking. It should be noted that in the majority of cases, albeit not all, any expression of intent to commit a clear criminal act was condoned by members of the community. However opinions and beliefs around the grey hat areas were more subtle and complex. There was a clear passion and skillset amongst the majority of users to make systems better and more secure. Harnessing this knowledge may help society address cybersecurity challenges.

References

1. Olenick, D. *Cybersecurity job gap grows to 3 million*. 2018 29/4/19; Available from: https://www.scmagazine.com/home/security-news/cybersecurity-job-gap-grows-to-3-million-report/.
2. Thackray, H., et al., *Surveying the hackers: The challenges of data collection from a secluded community* in *16th European Conference on Cyber Warfare and Security*. 2017: Dublin.
3. Holtz, P., N. Kronberger, and W. Wagner, Analyzing internet forums: A practical guide. Journal of Media Psychology: Theories, Methods, and Applications, 2012. 24(2): p. 55-66.
4. Pennebaker, J.W. and L.A. King, *Linguistic styles: Language use as an individual difference.* Journal of Personality and Social Psychology, 1999. **77**(6): p. 1296-1312.
5. Seebruck, R., *A typology of hackers: Classifying cyber malfeasance using a weighted arc circumplex model.* Digital Investigation, 2015. **14**: p. 36-45.