The Cubicle vs. The Coffee Shop: Behavioral Modes in Enterprise End-Users

Frédéric Giroire[1], Jaideep Chandrashekar[2], Gianluca Iannaccone[2], Konstantina Papagiannaki[2], Eve M. Schooler[2], and Nina Taft[2]

[1] INRIA, France
frederic.giroire@inria.fr
[2] INTEL Research
first.initial.last@intel.com

Abstract. Traditionally, user traffic profiling is performed by analyzing traffic traces collected on behalf of the user at aggregation points located in the middle of the network. However, the modern enterprise network has a highly mobile population that frequently moves in and out of its physical perimeter. Thus an in-the-network monitor is unlikely to capture full user activity traces when users move outside the enterprise perimeter. The distinct environments, such as the cubicle and the coffee shop (among others), that users visit, may each pose different constraints and lead to varied behavioral modes. It is thus important to ask: is the profile of a user constructed in one environment representative of the same user in another environment?
In this paper, we answer in the negative for the mobile population of an enterprise. Using real corporate traces collected at nearly 400 end-hosts for approximately 5 weeks, we study how end-host usage differs across three environments: inside the enterprise, outside the enterprise but using a VPN, and entirely outside the enterprise network. Within these environments, we examine three types of features: (i) environment lifetimes, (ii) relative usage statistics of network services, and (iii) outlier detection thresholds as used for anomaly detection. We find significant diversity in end-host behavior across environments for many features, thus indicating that profiles computed for a user in one environment yield inaccurate representations of the same user in a different environment.

1 Introduction

Traditional studies of end-user behavior in a network typically have employed traffic traces collected from network aggregation points (routers, switches, firewalls, etc.). In modern enterprise networks, a large subpopulation is mobile; laptop users move seamlessly in and out of the corporate office daily. When outside, the end-hosts are used in a number of places such as homes, airport lounges, coffee shops, etc. The VPN infrastructure of the enterprise ensures that users are never really cut-off from the resources on the corporate LAN. In fact, with the growing trend to support flexible telecommuting policies, and the ubiquity of network connectivity while outside the corporate network, users spend fewer

hours physically within the office cubicle, or at least within a single work locale.

Usage models are quite different inside and outside the office for a variety of reasons. Infrastructure services (email, directory, and print services) may simply be unavailable when users are outside the enterprise. Furthermore, locations outside the enterprise often have noticeable resource limitations (less bandwidth, less security, et cetera). Thus, users may be hampered from listening to streaming music, or may be wary of checking bank accounts when in a coffee shop. Conversely, the corporate acceptable usage policy may prohibit peer-to-peer file sharing applications on the corporate LAN, whereas it may be a staple application at home.

Previous work on building user-based profiles, such as in [1–5], does not consider the modality of the end-host when it is outside the enterprise. We argue that the growing trend to work outside the office and the distinct "usage-models" across the different environments, renders the single-view profile of the end-host (like the one generated from enterprise measurements alone) incomplete. In this paper, we explore the hypothesis that a single (static) profile for an end-host is inconsistent and/or incomplete. This has important consequences across the domains of enterprise security, network design, capacity planning and provisioning.

We analyze detailed traffic traces from a real corporate enterprise, where the traces were collected on the end-hosts themselves. This is in stark contrast to previous enterprise studies based on aggregate traffic, such as in [6, 7]. With these traces, we quantify the differences in behavior of the individual end-hosts across three different environments in which they operate: (i) inside the corporate enterprise, (ii) outside but connected through the corporate VPN, and (iii) outside, meaning disconnected from the enterprise altogether. To the best of our knowledge, this dataset is the first to capture traffic at end-hosts themselves. By collecting traces in-situ, rather than in network, we are able to correctly track a host's traffic even when its address, location, and/or network interface changes - avoiding the difficulties posed by DHCP address changes and host mobility that can thwart the accuracy of in-network traffic traces.

In this initial exploration of the "environment diversity" hypothesis, we focus on three distinct types of features. These are (i) the median duration of a user's presence in each environment, (ii) the relative usage of network services (destination IP ports) per environment for end-hosts, and (iii) outlier detection thresholds (the 95^{th} percentile) for TCP/UDP/ICMP connection counts as used by anomaly detection.

The contributions in this paper improve and clarify our understanding of end-host user profiles. Although our central hypothesis, i.e., that profiles need to change across environments, seems obvious, there has been no previous research quantifying such a hypothesis. This is most likely due to lack of availability of the right kind of data for such a study. This paper aims to explore this gap in end-host traffic characterization.

2 Data Description

Our dataset consists of packet traces collected at nearly 400 enterprise end-hosts (5% desktops and the rest, laptops) spanning approximately 5

weeks. A novel aspect to these traces is that they were collected *on* the individual end-hosts; this provides visibility into the end-host's traffic even as it leaves the office environment. Participants in our data trace collection were geographically distributed; 73% of the users were from the United States, 13% from Asia, 11% from Europe, less than 1% in each of Israel, Ireland and Latin America. All but a few users were based out of large offices in metropolitan areas. All the hosts in the study ran a corporate standard build of Windows XP. We solicited employees to sign up on a voluntary basis for the trace collection via organizational mailing lists, newsletters, and so forth. Cash prizes were offered as an added incentive to participate. Participants explicitly downloaded and installed the data collection software on their personal machines, thereby giving consent. We estimate that approximately 4000 employees were solicited, out of which approximately 1 in 10 installed the software. Overall, the data collection effort yielded approximately 400 GB of traces.

The collection software was written as a wrapper around the `windump` tool that logs packets in the well-known `pcap` format. The wrapper tracked changes in IP address, interface, or environment; upon such a change, `windump` was restarted and a new tracefile created. Importantly, every trace file was annotated with flags indicating the active network interface, the environment and if the logical VPN interface was active. Once installed, the software ran continuously (when the machine was on) for 5 weeks. For some users, it ran a few days less as they did not install the software immediately. Corporate policy strongly discourages the use of P2P applications, and hence our set of users is unlikely to be using any such software, even when outside the corporate environment.

To mitigate privacy concerns, we only collected the first 150 bytes of each packet. We did this simply to be able to infer the actual external destination when the packets went through the corporate proxy server. After identifying the actual destination, the payloads were discarded and only the packet headers retained. The post processing was carried out on a central server where traces were periodically uploaded. Moreover, all naming information regarding the user identity or machine identity was discarded upon upload of the traces. All solicitation emails contained a complete description of the data to be collected, the anonymizing procedures, and a disclosure of how the data was intended to be used. Because of this anonymization, we cannot know which traces came from engineers, managers, executives, etc.

Importantly, *all* the end-hosts in the study were *personally issued*, i.e., there is a single user per host. This is because in our corporation, each employee is given one laptop as their primary computer. Some employees, as needed, are additionally issued desktops; these are primarily used for running tests, simulations, etc. Most employees take their laptops home with them in the evening. Based on anecdotal evidence, employees generally shy away from allowing family members or others to use their computers. Hence we expect that the majority of our end-hosts have a single user, even when outside the corporate environment. Although a single user may use multiple machines, our intent here is *not* to characterize all aspects of the user at all times. Instead the focus is on all aspects of how a user uses a particular machine. This is what impacts whether

or not a single machine should switch profiles as it, together with a user, moves between environments. In that sense, it does not matter what the user does with other machines.

3 Diversity Across Environments

Users move between three different environments– that we call `inside`, `vpn`, and `outside`. In the first, `inside` (the corporate network), the end-host is plugged into the office LAN almost always with a wired ethernet connection (on occasion connecting to the wireless LAN). In our enterprise, employees use laptops as their primary computer system, and while at work, these move between a docking station (at their desk), meeting rooms, corporate cafeterias, etc. In the `vpn` environment, users launch a VPN client that "logically" connects them into the office LAN. Note that here, users could be outside the office (the common case), or inside where they are on an unsecured wireless network, which exists solely as a gateway to the VPN, and cannot be used to reach the outside. Finally, when `outside`, the user is physically outside the enterprise network, and does not have any access to any of the enterprise infrastructure services (email, file & print server, etc.).

As an initial glance into our data, we show the movements of two users between these environments. In Fig. 1, we show a three week timeline. Here, the width of the contiguous blocks denote occupancy in that environment. First, we observe that both users actually use all three distinct environments. Although not shown here, due to lack of space, this is true for the vast majority of users (there were very few exceptions). Second, we note that these two users have very different behaviors in terms of how much time they spend in each environment, and how frequently they switch between environments. The user on the right is primarily in `vpn`, indicating that he may travel considerably or work from home. This user also tends to leave his VPN connection open during much of the weekend. This could indicate one of two things, either our user is someone who wants to be able to respond quickly when email arrives, or someone who perceives (as is common) that his computer is safer when the VPN is active. In contrast, the user on the left seems to have a more traditional work and leisure time pattern, using the `inside` mode during daytime on weekdays, the `vpn` mode in the evening on weekdays, and the `outside` mode on weekends. Clearly the `outside` mode for this user is likely to capture non-office related activities.

What is obvious here is that different users have different needs, at different times, to access the resources on the enterprise network. Aside from diversity across users, it is also natural to expect that a single user carries out different activities in the different modes. We now explore such behavior for a variety of measures.

3.1 Environment Lifetimes

Motivated by Fig. 1, we first ask how much time a user spends in each environment. We define *environment lifetime* as the duration of contiguous time a user spends in a particular environment before changing it,

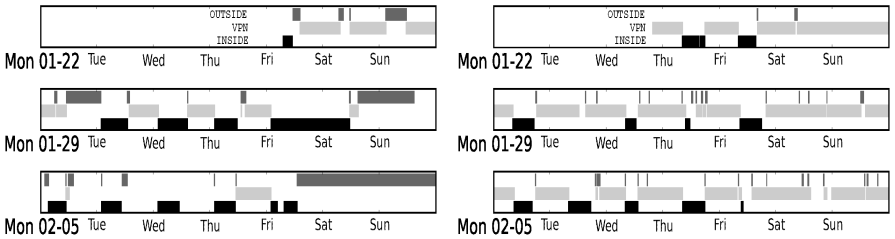
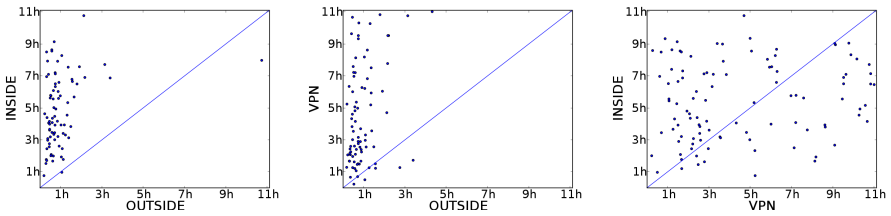


Fig. 1. A Tale of Two Users: time-line of two end-hosts over a 3 week window in the trace collection period.

restarting the machine or making it hibernate. Studying this statistic is key to solving many network design and planning problems. For instance, if one could model the time spent by users logged onto the VPN, the network operators could provision the VPN lines efficiently.

Fig. 2 is a set of scatter plots: each of these plots the median environment lifetime for individual users for two environments. In figure 2(a), each (x, y) point corresponds to a single user: the x value is the median time for `inside`, and y corresponds to `outside`. Similarly, Fig. 2(b) compares the lifetimes over `outside` and `vpn`, and finally, Fig. 2(c) compares `vpn` with `inside`. From these figures, it is quite clear that there is a marked difference in how long, in a single sittting, a user stays in each of these environments. Not surprisingly, for the most part, users spend more time `inside` as compared to the other two modes. It is interesting to see how short the environment lifetimes typically are for the `outside` mode. The lifetime spent `outside` can be anywhere from half to 10 times less than the typical lifetime for either the `inside` or `vpn` modes. An intuitive explanation for this could be that (i) the natural workday itself constitutes a window in which the employee is likely to stay in a single mode, and in addition (ii) when outside of work, the user's attention span (and time) is likely to be partitioned across mornings, evening, weekends, and interrupted by other domestic activities (meals, kids, etc.) which lead to shorter durations spent contiguously in the `outside` mode.



(a) Avg. Diff: +230% (b) Avg. Diff: +200% (c) Avg. Diff: +85%

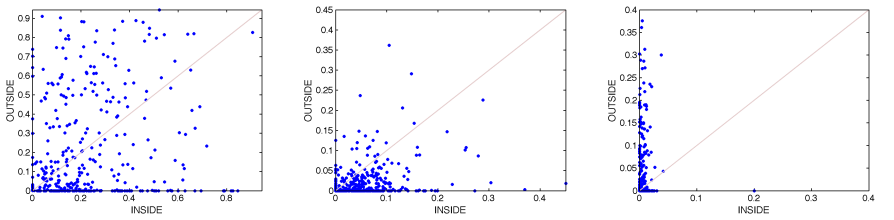
Fig. 2. Median Lifetimes in different environments. Median values across users: `outside`=43min, `vpn`=3h45min, `inside`=6h30min.

In comparing the environment lifetimes of `inside` mode versus `vpn` mode, we find interestingly that users exhibit tremendous diversity: some can stay on the `vpn` for 3 to 4 times as long as `inside`; others illustrate exactly the opposite behavior (points spread equally on both sides of the diagonal in Fig. 2(c). Users whose points lie near the extreme right side of this plot are likely to be employees who travel frequently, or who telecommute often, and thus their dominant work environment is through a VPN. We also observe, that even within the `inside` mode, there is great diversity across users - some have working sessions for 8 to 9 hours, while for others the median time is 1 or 2 hours. The main takeaway from these statistics is that we see two kinds of diversity. There is tremendous diversity for each individual, in terms of the time the user stays pinned to particular environments. Not only do users spend vastly different amounts of time in each environment, but knowing a particular user's behavior does not reveal much about the others. Some users will have similar trends (regarding the fraction of time spent in each environment), whereas others exhibit completely opposite trends. We thus also see diversity across users for this measure.

3.2 Destination Port Diversity

We now examine whether there are quantitative differences in how network services are used in different environments. We use TCP and UDP destination ports as a useful proxy for "network service" (for the subset of ports we consider this is reasonable). Because it is impossible to exhaustively examine all destination ports, we focus on two logically formed groups. First, we study the ports associated with HTTP and web traffic (80,88,8080, 443) which we term the Web ports, and second, we look at the ports associated with Windows based services, that are popular in the enterprise (135,389,445,1025-1029), denoted MS Ports.
The particular metric of comparison that we use is the fraction of connections corresponding to a particular port (or group of ports). For every user and in each of the environments, we collect all the connections made to a particular port and the metric is computed as the *ratio of connections on that port to the total number of connections* (in the same environment). This is intended to capture a notion of what percent of a user's activities in each environment do they spend on a given service. Fig. 3 plots this metric for three different port sets, in each case comparing behavior across the `inside` and `outside` modes. In each scatter plot, a point corresponds to an individual user and the (x,y) coordinates are the connection fractions corresponding to `inside` and `outside`, respectively. Fig. 3(a) plots the statistic for http traffic across the `inside` and `outside` environments (we exclude SSL traffic on port 443 from this plot and analyze that separately). The first thing we observe is the scattering of points over the entire graph. Importantly, nearly all these points are off the diagonal, indicating the percent of activity spent browsing the web in the two environments is not the same for users. Interestingly, there is no "typical user". For some users, the fraction of connections they generate that are HTTP is higher when in `inside` mode, and for other users, it is the reverse. The set of dense points along the x-axis indicates users

that only use HTTP when inside the enterprise. Such users may have a second machine at home that they use for general browsing. Such users stand in contrast to the user at (0.1,0.8) who generates 8 times as many HTTP connections (as a function of his total traffic) when outside as opposed to when at work. This could capture a user that prefers to read news, or pursue other leisure activities, when outside the office.



(a) Web (HTTP) traffic (b) Web (SSL) traffic (c) Microsoft/Windows traffic

Fig. 3. Comparing behavior across `inside` and `outside` environments.

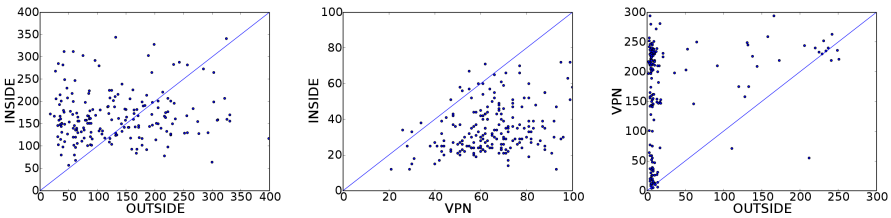
Similarly, we see in Fig. 3(b) for SSL traffic, that most of the points are off diagonal. Depending upon the user, the points can be a little or very far away from the diagonal. On most laptops, SSL constitutes a larger fraction of the total activity when the machine is inside the enterprise. In Fig. 3(c), we see a dramatic difference in the use of the MS ports. This is not surprising as many of these are primarily infrastructure services. The three plots confirm our hypotheses, that activity level profiles for a user are not the same in different environments. For some users, the differences may be small (but nonzero), whereas for others, the differences can be dramatic. We thus advocate that any profiling methodology that attempts to capture relative traffic measures — of a network service, for a given user on a particular machine — needs to be environment aware.

3.3 Thresholds on Behavioral Anomaly Detectors

Today, most enterprise end-hosts employ Host Intrusion Detection Systems (H-IDS) for security purposes. H-IDS systems typically include, among other things, a suite of anomaly detectors. From recent research, a popular approach to anomaly detection is to build behavioral profiles and use them to understand what is and isn't "normal" at an end-host. Many anomaly detectors define a threshold, [8–10], which defines the boundary between what is normal and abnormal for that host.

We now ask the important question as to whether or not such thresholds would vary for a given user, across different environments? If so, this would imply that the configuration of anomaly detectors also needs to be environmentally-aware, possibility loading different profiles (i.e., thresholds) into the H-IDS, depending upon the current user environment.

Some detectors track the number of connections of a particular type within a time window. Here we will examine this type of feature for



(a) TCP connections (b) UDP connections (c) ICMP traffic

Fig. 4. 95^{th} %-ile values for *tcp, udp* and *icmp* protocols. Connection counts in 15 minute windows are used.

TCP connections, UDP connections and ICMP packet-pair counts. For these 3 protocols, we count the number of connections in 15 minute windows and build histograms for each user to indicate how many are likely. We compute thresholds that demarcate the 95^{th} percentile point of these distributions, and consider these as the threshold value for the anomaly indicator. Considerable work has been devoted to the very specific problem of selecting suitable definitions of what constitutes an anomaly, or an outlier, however this topic is well outside the scope of this paper. Instead we pick a simple definition of an outlier and use it consistently across users and environments; this facilitates a straightforward comparison of the tail behavior of users across environments. Here, we study how this value differs across the environments.

In order to obtain connection records from raw packet traces, we use `bro` [11] to reassemble the flows/connections from the packet headers. The 95^{th} percentile values for the three features are shown in Fig. 4. In each scatter plot, a point corresponds to the values, in the two environments, for an individual user.

The high-level observation is that points are considerably off diagonal in every case. Note that points on or near the diagonal correspond to users that have approximately the same threshold value in both environments being compared. For instance, take fig. 4(a): here most of the points are well off diagonal. Moreover, roughly half of the user population lies on either side of the diagonal. The latter hints at two user classes (of roughly equal population) for whom the value in one environment dominates. Take the point most extreme to the right: the 95%-ile corresponding to `outside` is 400, and 120 for `inside`. Thus, there is a higher intensity of outgoing TCP connections when in `outside`, while when in `inside`mode there are almost no 15 minute windows in which one sees more than 120 TCP connections. This is a marked difference. If a security anomaly detector tracking TCP connections, is configured with a threshold of 120/15min for all environments, then when the machine is `outside` a large number of false positives will be generated. Conversely, if the machine were configured with 400 connections/15min, then when the machine was in `inside`mode, it would miss all stealthy attacks. Clearly neither of these is good for all environments.

Fig. 4(b) plots the differences for UDP flows; here, we contrast usage in `vpn` and `inside`. We clearly see that the bulk of the distribution is away from (and below) the diagonal; this signifies that one sees more UDP flows in `vpn`, as opposed to `inside`. This seems puzzling at first; one would normally expect more traffic, and correspondingly larger number of UDP flows when `inside`. Upon closer inspection, we identified two destination UDP ports that contributed a large number of small sized flows; one was associated with the VPN client application and the other with a software compliance checker. The flows from these ports contributed significantly to the "rightward" skew of the points in Fig. 4(b). When the same plot was recomputed after filtering out flows from these two ports, the distribution of the points more closely resembled that in Fig. 4(a).
Finally, in Fig. 4(c) we compare ICMP traffic across `outside` and `vpn` environments. We see that there is very little ICMP traffic (to almost none) when the host is `outside`. Thus, ICMP traffic is extremely discriminating to the environment (more than TCP and UDP). This is possibly due to a lot of maintenance and network management traffic when the machine is on the VPN (a logical extension of the enterprise). This last observation strongly supports our hypothesis, i.e., that environment awareness is critical. A number of DDoS attacks, and some OS fingerprinting techniques make use of ICMP probes; large amounts of ICMP traffic are generally suspicious. In the figure, we see many users generate 200-300 ICMP packets within 15 minutes, and to be effective, an anomaly detection threshold would be set above this level. However, when we do this, we essentially provide a safe margin of the same amount (of ICMP traffic) when the host is `outside`; an infected or compromised machine could send out 200-300 ICMP packets without any fear of being flagged.
We conclude from this section, that because thresholds used by anomaly detectors define a boundary between normal and abnormal traffic, end-user based security mechanisms need to be designed to be "environment-aware". This is because these boundaries **do** change across environments for the same user.

4 Conclusion

Our study of common user-behavior features illustrates that most users exhibit significant diversity in how they use their machines in different environments. We show this on traces collected from end-hosts in an actual enterprise network. Regardless of whether we are looking at time spent in an environment, volumes of connections, http traffic, fraction of connections for Microsoft/Windows services, the measure can differ by anywhere from twice to 10 times as much in one environment as compared to another. These results illustrate that a profile computed in one environment will yield an inaccurate representation of user activity levels in another environment, for the majority of the users.
We showed how this could impact the configuration of anomaly detectors in H-IDS systems. These findings have implications for a number of other applications as well, such as resource allocation, VPN tunneling, and even virtual machine configurations. For example, if tomorrow's

laptops employ different virtual machines for the home and work environments (such as the "red/green" VM proposal in [12]), then each VM should be configured to grab the appropriate profile before launching. We thus believe that "environmental awareness" is important for such applications. In the future we plan to study how some of these applications could be improved by using environmentally-aware profile information. We also plan to carry out user clustering to determine the minimal number of common profiles that could be used to capture the entire set of user behaviors.

References

1. McDaniel, P., Sen, S., Spatscheck, O., der Merwe, J.V., Aiello, B., Kalmanek, C.: Enterprise security: A community of interest based approach. In: Proc. of Network and Distributed System Security (NDSS). (Feburary 2006)
2. Tan, G., Poletto, M., Guttag, J., Kaashoek, F.: Role classification of hosts within enterprise networks based on connection patterns. In: Proc. of the USENIX Annual Technical Conference 2003, USENIX (2003) 2–2
3. Karagiannis, T., Papagiannaki, K., Taft, N., Faloutsos, M.: Profiling the end host. In: Passive and Active Measurement. (2007) 186–196
4. Padmanabhan, V.N., Ramabhadran, S., Padhye, J.: Netprofiler: Profiling wide-area networks using peer cooperation. In: IPTPS. (2005) 80–92
5. Bhatti, N., Bouch, A., Kuchinsky, A.: Integrating user-perceived quality into web server design. In: Proc. of the 9th International World Wide Web conference on Computer networks, North-Holland Publishing Co. (2000) 1–16
6. Pang, R., Allman, M., Bennett, M., Lee, J., Paxson, V., Tierney, B.: A first look at modern enterprise traffic. In: Proc. of the Internet Measurement Conference (IMC), ACM (2005) 2–2
7. Bahl, P., Chandra, R., Greenberg, A., Kandula, S., Maltz, D.A., Zhang, M.: Towards highly reliable enterprise network services via inference of multi-level dependencies. In: Proc. of ACM SIGCOMM, New York, NY, USA, ACM (2007) 13–24
8. Biles, S.: Detecting the unknown with snort and the statistical packet anomaly detection engine (SPADE) Computer Security Online Ltd.
9. Jung, J., Paxson, V., Berger, A.W., Balakrishnan, H.: Fast portscan detection using sequential hypothesis testing. IEEE Symposium on Security and Privacy (2004) 211
10. C. Kreibich and A. Warfield and J. Crowcroft and S. Hand and I. Pratt: Using Packet Symmetry to Curtail Malicious Traffic. Fourth Workshop on Hot Topics in Networks (HotNets-IV) (November 2005)
11. Paxson, V.: Bro: A system for detecting network intruders in real-time. COMPUT. NETWORKS **31**(23) (1999) 2435–2463
12. England, P., Manferdelli, J.: Virtual machines for enterprise desktop security. Information Security Technical Report **11**(4) (2006) 193–202