From: [redacted] </O=ITHAKA/OU=FIRST ADMINISTRATIVE GROUP/CN=RECIPIENTS/CN=[redacted]>
Sent: Sunday, October 10, 2010 10:52 AM
To: [redacted]@ithaka.org>; [redacted] <[redacted]@ithaka.org>
Cc: [redacted]@ithaka.org>
Subject: RE: MIT is currently denied JSTOR Access

Thanks [redacted]

Just to clarify at the outset, the 1st incident that happened 2 weeks ago (9/25 – 9/26) was the same incident they reported a visiting scholar responsible for, the 2nd incident started yesterday. I was mistaken, It was not 3 weeks ago, it was 2 weeks.

The 1st incident was excessive, but seen as one off and part of the normal abuse protocols. Robot being used, IP access being denied, institution being responsive and asserting they had taken care of it. We have any number of these on a weekly basis, but to the tune of a few hundred PDFs per case. This was to the tune of several thousand from what I understand.

I recognize that this is not an ordinary case and I should have looped you in on it at the time, that was poor judgment on my part. I'll be sure to do so going forward. At the time, [redacted] was satisfied with the response, as was I, and the activity did not recur. It was seen as an efficient robot, but not special, just using a new patter and with access to fantastic bandwidth through MIT. Definitely not any kind of denial of service attack.

Basically, when you apply limits based on session as we have, the natural progression is to adjust the pattern, one of which is to simply start new sessions more frequently.

[redacted] and I discussed this very pattern last year when abuse protocols were first put in place. The only weapon we have against it, based on the pattern, is some version of CAPTCHA, requiring entry for the 1st download of every session and once again, randomly, every N downloads. This would prevent this particular kind of scripting. My opinion on that is that if these are isolated incidents, that barrier is too high for the regular user. [redacted] may not feel that way. From my perspective MIT is a special case and the first of this kind of scripting we have seen on any scale and the first to threaten the live site with its efficiency. They will likely be embarrassed that this has recurred.

It is also not clear to me if they were scraping the site heavily for metadata, stable URLs, downloading PDFs or all of the above. We are adjusting the abuse tools down to try and auto-prevent, but we need precise data in order to both prevent this kind of abuse pattern and not disrupt normal use elsewhere. To that end I will ask [redacted] and [redacted] to provide a recap of the incident and the numbers surrounding it as they never hit our abuse monitoring rules the 1st or 2nd time, I have no window into the numbers themselves.

The second occurrence, as it stands, started at some point yesterday, but I am unclear as to when that was. The [redacted] summary will give us that.

As for the timeline to this point, [redacted] began communicating broadly about it shortly before 6 pm yesterday and the IP range was denied around 6:45 pm. I first saw the email chain from [redacted] around 8:30 pm and began communicating to MIT, [redacted] around 9 pm. MIT has not responded and the IPs will remain blocked until they do.

I will be sure to keep you both looped in as progress is made against both resolving this issue and providing a summary report.

[redacted]

From: [redacted]
Sent: Sunday, October 10, 2010 9:33 AM
To: [redacted]
Subject: Re: MIT is currently denied JSTOR Access

Thanks, [redacted] Very troubling.

Also, this is the first I've heard of this. [redacted] needs to be involved earlier in these types of matters to ensure we adhere to contract terms and pursue appropriate remedies. This is part of the protocol we established. The activity noted is outright theft and may merit a call with university counsel, and even the local police, to ensure not only that the activity has stopped but that -- e.g. the visiting scholar who left -- isn't leaving with a hard drive containing our database.

[redacted], please clarify the history of this activity. What happened 3 weeks ago? How did we respond? What has happened in the intervening time?

And what is the "last time" incident with the visiting scholar? Is this the one from three weeks ago? Again, this is the first I've heard of this. Going forward, we need to make sure [redacted] is aware of these cases as they are emerging. The [redacted] component is critical here, but there are other avenues that often need to be pursued concurrently.

[redacted]

From: [redacted]
Sent: Saturday, October 09, 2010 11:18 PM
To: [redacted]
Subject: Fw: MIT is currently denied JSTOR Access

Just want to be sure [redacted] is aware of the situation.

[redacted]

From: [redacted]
Sent: Saturday, October 09, 2010 10:30 PM
To: [redacted]
Cc: [redacted]
Subject: MIT is currently denied JSTOR Access

Good Evening,

I want to make you aware that the MIT abuse case that showed up 3 weeks ago came back today, forcing [redacted] to deny 18.0.0.0/8 at the firewall. Half of Manchester needed restarting this afternoon to address servers that got jammed up due to this activity. I have emailed our contacts at MIT and informed them of the situation.

Just for clarity, that's the whole range, the largest we have ever denied. Last time they reported that a visiting scholar was responsible and that it should not recur as the scholar had left.

I am in conversation with [redacted] and [redacted] is reporting that this scraping is very intensive and threatening the website when unblocked. The block of their range has brought the incident under control and they are currently getting deny pages and not threatening the website.

The pattern is simple… they start a session, download 1 pdf, start a new session, download 1 pdf on and on. [redacted] can comment on the specific volume and duration.

Also, after the last incident at MIT, we implemented Literatum's # of sessions per hour IP blocking rule to 5000 sessions in 60 minutes. It did not fire. We are digging deeper, but the earliest speculation is that this rule is applied 'per server', which we did not anticipate, meaning we'll need to adjust the number down based on data gathered from this incident and elsewhere.

I don't know if this will cause any negative reaction in the public and haven't heard anything through our feedback channels as yet, but wanted to make sure we were all on the same page and that there are no surprises here. This is an extreme block to combat an extreme attack.

More as the situation gets resolved.

Best,

[redacted]
[redacted]
JSTOR | Portico

[redacted]@ithaka.org
[redacted]