(12:09:44) David Gucwa: k
(12:45:15) David Gucwa: There's a hitch in the plan to crawl through the profiles in advance
(12:45:43) David Gucwa: when logged in, I can only seem to access profiles that are on my friends list or at the same school as I am
(12:47:17) David Gucwa: so I think we're going to have to spider them on the spot when people import
(12:47:24) David Gucwa: and cache them then
(12:59:32) dr ttol: wait
(12:59:42) dr ttol: atleast we can index entire schools
(12:59:43) dr ttol: with just one email
(12:59:44) dr ttol: right
(12:59:52) dr ttol: so we just need to get you 300 logins
(12:59:55) dr ttol: or so
(13:00:03) dr ttol: to access over 1 mil profiles
(13:00:48) David Gucwa: yeah
(13:01:03) David Gucwa: if you can just give me one big file with all the logins and passwords then I can use that
(13:05:07) dr ttol: can you use the one we gave you
(13:06:12) David Gucwa: I'm not sure this is going to work actually
(13:06:32) David Gucwa: I just timed the script, and it takes about 5 seconds to fetch a profile
(13:06:44) David Gucwa: for a million profiles that's 57 days to index them all
(13:06:48) dr ttol: thats fine
(13:07:33) dr ttol: we just want to get as much as we can
(13:07:47) dr ttol: i mean, if we have all of umass
(13:07:51) dr ttol: thats great
(13:07:57) dr ttol: prob only take a few hours
(13:08:33) David Gucwa: well, except I don't just have a list of ids that belong to umass
(13:08:36) David Gucwa: I have to try each one
(13:08:52) dr ttol: what do you mean
(13:09:34) dr ttol: with one umass login and password
(13:09:38) dr ttol: you're able to see the entire school
(13:09:57) David Gucwa: If I type in the id of a profile, and that person is from umass, then I can see it
(13:10:05) David Gucwa: but there's no way to know which ids are from umass
(13:10:14) dr ttol: if you click browse
(13:10:21) dr ttol: you get a list of users
(13:10:30) dr ttol: that are from you school
(13:11:17) David Gucwa: where does it say browse
(13:11:27) dr ttol: give me a login and password
(13:11:51) David Gucwa: lbowman@wellesley.edu divya
(13:12:52) dr ttol: social net
(13:12:56) dr ttol: is browse
(13:12:58) dr ttol: it is randomly generated
(13:13:01) dr ttol: so you have to check for duplicates
(13:14:39) dr ttol: see it?
(13:15:05) David Gucwa: yeah
(13:15:09) David Gucwa: let me see how fast I can grab ids from that
(13:15:13) dr ttol: k
(13:15:18) dr ttol: and that "next" link
(13:16:06) David Gucwa: next is just the same as reloading
(13:16:38) dr ttol: k
(13:20:10) David Gucwa: I can get 3 or 4 ids a second
(13:21:07) David Gucwa: How many people from umass would you say are on thefacebook? A few thousand?

(13:21:56) David Gucwa: It'll probably take ~6 hours to index all of umass's profiles
(13:22:03) David Gucwa: and cache them
(13:24:11) David Gucwa: actually that's not true, since I can only get a random list of them
(13:24:42) David Gucwa: I can probably get a large chunk of the ids pretty quickly, but towards the end it's going to be harder to get ids I haven't seen yet
(13:25:04) David Gucwa: when does this project have to be finished?
(13:28:15) dr ttol: which project
(13:28:29) David Gucwa: the whole thing, when is it going to go live
(13:28:46) dr ttol: later this week, so we want the spider to continue to do as much as possibl
(13:29:03) David Gucwa: Yeah the spider is going to be doing the bulk of the indexing
(13:29:18) dr ttol: would two spiders indexing at the same time be twice as fast
(13:29:30) dr ttol: would three? would X?
(13:30:11) David Gucwa: if we put them all on different servers. I'm not sure how much of a bottleneck bandwidth or cpu power would be if we had X spiders running on one computer
(13:30:46) dr ttol: and one central index
(13:31:47) David Gucwa: how many proxies do we have access to at the moment
(13:31:54) dr ttol: a few
(13:31:59) dr ttol: ill order more every day
(13:32:32) David Gucwa: well I'll index what I can
(13:32:43) dr ttol: ok
(13:32:55) David Gucwa: do you have a umass login that I can use
(13:33:52) dr ttol: one sec
(13:37:58) dr ttol: rdegutis@student.umass.edu
(13:38:02) dr ttol: ruth1783
(13:41:28) dr ttol: we can have all of umass indexed in 6 hours?
(13:47:31) dr ttol: ?
(13:49:46) David Gucwa: I can have a lot of it indexed in 6 hours
(13:49:57) David Gucwa: the problem is the randomness
(13:50:13) David Gucwa: like once we have 99.9% of the id's, then we're going to almost always get repeats
(13:50:22) David Gucwa: the more we have the harder it is to get new ones
(13:50:32) dr ttol: right
(13:50:39) David Gucwa: it's diminishing returns
(13:50:44) dr ttol: so get as much as you can, then move into another school
(13:50:47) David Gucwa: k
(13:50:51) dr ttol: how many ids in 6 hours
(13:51:06) David Gucwa: At least a thousand
(13:51:29) dr ttol: tomorrow, can we set it up so that we can crawl two separate schools at the same time
(13:52:07) dr ttol: does it take 5 seconds from your house?
(13:52:25) dr ttol: would it be faster if it was on i2hub's web machine
(13:52:56) David Gucwa: probably
(13:53:12) dr ttol: can you test the difference
(13:56:02) David Gucwa: yeah
(14:03:36) David Gucwa: I'm not sure this is going to work actually
(14:03:39) dr ttol: ?
(14:03:59) David Gucwa: I'm running into problems getting the social network page to show from commandline
(14:04:59) dr ttol: well, they are reachable by browser
(14:05:06) dr ttol: so its not impossible

(14:06:32) David Gucwa: the thing about doing it from php is that it doesn't remember that you're logged in, so you can't go to the login page and then go to the search page
(14:06:45) David Gucwa: you have to go right to the search page and encode the login information into the url
(14:07:04) David Gucwa: thefacebook doesn't seem to like that for the search page for some reason
(14:08:27) dr ttol: whats a solution
(14:08:33) David Gucwa: i'm looking for one
(14:11:07) David Gucwa: you mentioned that when importing, you want to be able to send invite emails to people who aren't in connectu yet
(14:11:25) dr ttol: yes
(14:11:44) David Gucwa: I'm not sure how vital that is, but if we don't add that functionality then we will not need to crawl into other people's profiles and importing will take 5 seconds instead of 5 minutes
(14:12:18) David Gucwa: we can still get a list of their friends by id and find those friends on connectu if they have profiles already
(14:12:23) dr ttol: well
(14:12:25) dr ttol: if we crawl now
(14:12:31) dr ttol: then we'll also know who they're friends with
(14:13:01) David Gucwa: I think it would be a lot less trouble to do it at import time
(14:13:06) dr ttol: so if we have software that can crawl through every profile on a school, we'll know all the social networks
(14:14:04) dr ttol: there are two approaches to this
(14:14:09) dr ttol: we can either:
(14:14:20) dr ttol: a) index through "social net"
(14:14:37) dr ttol: b) index at import time, and import allt heir friends and all the profile information of themselves and their friends
(14:14:44) dr ttol: both will likely have the same outcome -- an index of the entire net
(14:14:51) David Gucwa: I'm not sure why we need the profile information for their friends
(14:14:53) David Gucwa: at import time
(14:15:54) dr ttol: because when that friend signs up, we already have their profile information
(14:16:53) David Gucwa: we can go fetch it pretty quickly when that friend signs up
(14:18:05) dr ttol: we still need the email addresses
(14:19:02) David Gucwa: That's what I was asking, how important is that functionality
(14:19:10) dr ttol: pretty important
(14:19:38) David Gucwa: which is more important, a short import time or sending email invites to non-users
(14:19:52) dr ttol: cant we send the email invites later
(14:20:06) dr ttol: it'll still be short import time, we can just have another software do the crawling of non-users
(14:20:18) dr ttol: we want to be able to send friend requests for email addresses already in our database
(14:20:20) David Gucwa: is the invite coming from connectu or from the person who just signed up
(14:20:30) dr ttol: connectu
(14:20:36) David Gucwa: okay that's fine then
(14:20:46) dr ttol: but via importer software
(14:20:53) dr ttol: like a sendmail()
(14:21:12) David Gucwa: right
(14:21:25) dr ttol: but we still need to crawl