PRINCETON UNIVERSITY

Center for Information Technology Policy
3rd Floor, Sherrerd Hall
Princeton, New Jersey 08544
Email: citp@princeton.edu
T: 609-258-9658 F: 609-964-1855

March 30, 2011

The Honorable Lee H. Rosenthal
Chair, Committee on Rules of Practice and Procedure
Judicial Conference of the United States
Washington, D.C. 20544

Dear Judge Rosenthal:

I am a graduate student in computer science at Princeton's Center for Information Technology Policy. I research privacy and open government under the guidance of Prof. Ed Felten. I have been conducting research concerning privacy and redaction in electronic court documents, and I write to alert you to some preliminary results that may be of interest. Specifically, I have discovered almost two hundred documents from PACER containing confidential information that is hidden from view but can still be extracted with the right software. There are likely to be tens of thousands more documents like them in the PACER system.

In 2009, my colleagues and I created RECAP, software that enables users of the PACER system to automatically donate the documents they purchase to our free, public repository. Since the beginning of our project, we have been concerned with the potential harms of re-publishing inadequately redacted documents. And so I have been conducting research on ways to use the techniques of computer science to identify such documents so that they can be redacted or removed from our repository.

My current focus is on "bad redactions": cases in which information has been obscured by black rectangles, but can still be extracted with the right software. For example, the information can be extracted by cutting and pasting the text into another document, or the sensitive information under the rectangles could be indexed and duplicated by search engines.

This is a long-standing and well-known problem, but as far as I know no one has tried to automate the process of finding such documents. I have written software to automatically detect redaction failures and used it to analyze the approximately 1.8 million documents we had in our repository when I began my research. I found 194 documents with "bad redactions."

The majority of the documents (about 130) appear be from commercial litigation, in which parties have unsuccessfully attempted to redact trade secrets such as sales figures and confidential product information. Other improperly redacted documents contain confidential medical information, addresses, and dates of birth. Still others contain the names of witnesses, jurors, plaintiffs, and one minor.

On October 24, 2008, you received a letter from Carl Malamud of Public.Resource.Org detailing 1669 PACER documents with unredacted Social Security numbers and other confidential information. It is worth noting that the majority of our documents came from the documents Carl examined in 2008, and he redacted the problematic documents he found before he gave them to us. That means that the 194 documents I have discovered are *in addition* to the problem documents he identified.

Attached, please find a listing of the problem documents we have discovered. I hope the courts will modify or remove these documents from PACER and notify downstream users of bulk PACER documents to do the same.

Based on these findings, I hope you will permit me to make three recommendations. First, the courts themselves should be using software to automatically detect this type of problem with documents in the PACER corpus. Since we examined only a small subset of the documents in PACER, it is reasonable to expect that we would find tens of thousands of additional documents with this type of redaction problem if we examined the full PACER corpus. It would also be ideal to make a scan for "bad redactions" (and other privacy problems) a standard step in the CM/ECF submission process, so that such documents are caught before they are even posted to PACER. I would be happy to provide court IT staff with my source code (which is written in the free Perl programming language) and help the court's technical staff to modify it to work on the court's servers.

Second, it would be extremely helpful for the courts to provide academic researchers with access to large, random samples of documents from PACER. Although I was able to do interesting research using documents donated to us by our users and partners, the usefulness of the results is limited by the fact that the documents we have do not constitute a statistically-valid random sample of the PACER corpus. This makes it difficult to extrapolate our findings to all PACER documents with statistical rigor. Giving researchers access to a random sample of PACER documents would not only allow us to do more research that helps the American public better understand the judiciary, but may also allow us to do research (such as privacy audits) that is directly useful to the administration of the courts themselves.

Finally, I would like to add my voice to the many calls from stakeholders for a formal process for reporting privacy problems in PACER documents. I am likely to uncover additional problematic documents in the course of my research. It would be helpful to have a standardized process or point of contact for reporting such problems. I agree with the many others who have suggested the appointment of a Chief Privacy Officer—a Best Practice followed by many large organizations in government and industry. It might also be helpful to add a "report a problem" page to the PACER website that allows users to report problems in a standardized format.

Sincerely,

Timothy B. Lee
Ph.D. Candidate, Princeton University

Cc: The Hon. James C. Duff, Mr. Peter C. McCabe

File name	Case number	Pages	Redacted Info
almd	[redacted]	4, 6, 11, 21, 24, 29, 31, 36	Juror names
azb.		14, 17	Trade Secret
cofc.		1	Taxpayer ID number
cofc.		5	Social Security # (SSN)
cofc.		2	SSN
cofc.		2	SSN
cofc.		5	SSN
mnd.		6	SSN and Date of Birth
cacd		32, 33, 37, 38, 40, 75, 76, 82, 86, 87, 88, 102, 103, 108, 109, 110	Trade Secret
cand		6, 10	Trade Secret
cand		20	Trade Secret
cand		5-7, 9-16	Trade Secret
cand		2-7	Trade Secret
cand		8-11, 14, 17, 18, 20, 26-28, 31, 33, 35	Trade Secret
cand		3, 4, 8, 9, 11-23, 25-37, 39-42	Trade Secret
cand		6-8, 10-13	Trade Secret
cand		5, 8	Trade Secret
cand		3-9	Trade Secret
cand		1-4	Trade Secret
cand		8, 10, 19, 20, 22	Trade Secret
cand		2, 5-12	Trade Secret
cand		2-4	Trade Secret
cand		2, 4-7	Trade Secret
cand		2-3	Trade Secret
cand		2, 5-9	Trade Secret
cand		2-3	Trade Secret
cand		2-3	Trade Secret
cand		2-3	Trade Secret
cand		2-3	Trade Secret
cand		7-8, 11, 16, 24	Trade Secret
cand		2-3, 5-6	Trade Secret
cand		2	Trade Secret
cand		6, 9, 13	Trade Secret
cand		2	Prisoner name and booking number
cand		10, 17	Trade Secret
cand		17, 19	Trade Secret
cand		3	Trade Secret
cand		11, 21	Trade Secret

cand	[redacted]	12	Trade Secret
cand		11	Trade Secret
cand		1, 3, 5-9,12-13, 15, 19	Addresses
cand		3	Trade Secret
cand		6-8	Trade Secret
cand		3	Trade Secret
cand		9, 17-21, 23, 25-28	Trade Secret
cand		3, 6-7, 9, 13-27	Trade Secret
cand		9, 17-21, 23, 25-28	Trade Secret
cand		3, 6-7, 9, 13-27	Trade Secret
cand		70, 72, 74-78, 81-82,84,88	Addresses
cand		2	Trade Secret
cand		5-7	Trade Secret
cand		4	Trade Secret
cand		2-10, 13-14, 17-24, 26-28	Attorney billing
cand		1-2	Attorney billing
casd		15	Trade Secret
casd		6, 8-9, 11-13	Trade Secret
casd		8-9, 11-13	Trade Secret
casd		3-4, 7, 9, 11-12	Trade Secret
casd		3-10	Trade Secret
dcd.		4	National Security?
dcd.		8	Minor's DOB
dcd.		2-3, 10-38, 40-54, 56-62	Personnel dispute
dcd.		2-21	Personnel dispute
dcd.		2-3, 10-38, 40-54, 56-62	Personnel dispute
dcd.		2-21	Personnel dispute
dcd.		2-4, 7-13, 21-30	Trade Secret
dcd.		1-2	Employee issue
dcd.		2, 8, 10-17	Serial #, Address
dcd.		3-5, 30-31, 39,	medical/personal
dcd.		4, 14, 24	Trade Secret
dcd.		4-5, 13	Trade Secret
dcd.		11-13, 23, 28	Trade Secret
dcd.		2, 4-7, 9	Address
dcd.		2-3	Address
dcd.		2-3	Address
ded.		6	Trade Secret
ded.		4-5	Trade Secret
ded.		2-4	Trade Secret
ded.		2-4	Trade Secret
ded.		2-6	Trade Secret
ded.		4-6	Trade Secret
ded.		4-5	Trade Secret
ded.		3-4	Trade Secret
ded.		4-5	Trade Secret

ded.	[redacted]	2-3	Trade Secret
ded.		2	Witness names
ded.		2	Witness names/Trade Secret
ded.		2-4	Trade Secret
ded.		2-4	Trade Secret
ded.		19, 25	Trade Secret
ded.		11, 13-21, 28-29, 31-32, 35-39	Trade Secret
ded.		11-12, 14-27, 29-31, 33-34, 36-46	Trade Secret
ded.		12-17, 35	Trade Secret
ded.		19	Trade Secret
ded.		16	Trade Secret
ded.		9, 13, 15-17	Trade Secret
ded.		6-8	Trade Secret
ded.		7-9, 11-12, 15	Trade Secret
ded.		2-4	Trade Secret
ded.		2-4	Trade Secret
ded.		3-4	Trade Secret
ded.		2-4	Trade Secret
ded.		3	Trade Secret
ded.		299, 302-306	Trade Secret
ded.		6, 9-14, 16-19	Trade Secret
ded.		2-3	Trade Secret
ded.		1-3	Trade Secret
ded.		2, 5-8, 14, 16-18	Trade Secret
ded.		2, 7-13, 16, 23-26, 28-29, 31-39, 41-42	Trade Secret
ded.		2	Trade Secret
ded.		2	Trade Secret
ded.		5-8, 11-14, 16-17	Trade Secret
ded.		28	Trade Secret
ded.		2	Trade Secret
ded.		2	Trade Secret
ded.		4-7	Trade Secret
flsd.		10-17	Trade Secret
ilnd.		4-5, 9, 12	Trade Secret
ilnd.		8-11, 16, 20	Trade Secret
ilnd.		13	Trade Secret
ilnd.		14, 17-18	Trade Secret
ilnd.		5-7, 9-11, 14	Trade Secret
mad		16, 21-22	Trade Secret
mad		20-23	Trade Secret/medical
mad		2-3, 6, 9-10	Trade Secret
mad		2	PACER login
mad		3	PACER login

mad	[redacted]	2	PACER login
mad		3	PACER login
mad		4-6	Trade Secret
mad		7-11	Trade Secret
mad		11	Trade Secret
mad		11	Trade Secret
mad		11	Trade Secret
mad		6	Trade Secret/financial privacy
mad		9-10	Name of special needs individual
mad		3, 6	Trade Secret
mad		7-8	Trade Secret
mad		9, 23-25, 33-35, 38	Trade Secret
mad		7, 10-11, 13-14, 20-21, 23-26, 30	Trade Secret
mad		1-2, 4-9	Trade Secret
mad		3-4	Trade Secret
mad		16	Trade Secret
mad		20	Trade Secret
mad		14	Trade Secret
mad		9, 12-13	Trade Secret
mad		5-9	Trade Secret
mad		2, 5, 7-8	Trade Secret
mad		2-4, 6-8, 10, 12-25, 27-28, 36, 44-51, 59, 63-65	Drug investigation
njd.		1, 5-7, 9-11, 13	Trade Secret
nysd		1	Account number
nysd		6-16, 19-21, 24, 27	Trade Secret
nysd		3	Address
nysd		3	Account number
nysd		5	Security
nysd		6-7, 9	Medical
nysd		6-7, 9	Medical
nysd		6-7, 9	Medical
nysd		6-7, 9	Medical
nysd		5-6, 12-13	Plaintiff names
nysd		2	Plaintiff names
nysd		6-7, 9	Medical
nysd		6-7, 9	Medical
nysd		6-7, 9	Medical
nysd		6-7, 9	Medical
nysd		6-7, 9	Medical
nysd		5-6, 12-13	Plaintiff names
nysd		2	Plaintiff names
nysd		6-7, 9	Medical

nysd.	[redacted]	2-3, 6-8	Trade Secret
nysd.		2-4	Trade Secret
nysd.		6-7, 9	Medical
nysd.		2	Address
nysd.		89, 128-129, 131-132	Unknown
nysd.		6	Account number
nysd.		1-4	Account number
nysd.		3	Account number
nysd.		6	Account number
nysd.		1, 8	Account number
nysd.		2, 46	Trade Secret
nysd.		10-20, 23, 43-45, 48	Trade Secret
nysd.		3	Trade Secret
ohnd		8-11, 13-17	Trade Secret
utd.5		2	Attorney-client
utd.5		3-4, 7	Attorney-client
utd.5		2, 4, 6	Attorney-client
vaed.		20-21, 26	Trade Secret
vaed.		16-17	Trade Secret
vaed.		2-3, 6-8, 12-34	Trade Secret
vaed.		2-3, 7-19, 21-30, 32-35	Trade Secret
vaed.		2-4, 7-9, 18-20, 22, 25, 29, 31, 33	Trade Secret
vaed.		4-5, 11	Trade Secret
vaed.		4, 17, 20, 22, 25, 27	Trade Secret
wawd		4-10, 13-15, 22-23, 25	Personal