Let’s imagine this scenario.

You work at a tech company,

surrounded by code,
systems, and…
logs.

These logs capture everything: errors, performance, user actions.

Now, you need to share them.

Maybe with another team.

Maybe with a research group.

Or maybe you're using ChatGPT to ask,

“Hey, what went wrong here?”

But suddenly…

Someone from IT
sends you a message:

“Are you sure this data is safe to share?”

“Shouldn’t we anonymize it first?”

Now, that’s where the problem starts!

Your first question might be:

“What things should I anonymize?”

And your second question might be:

"How should I find all these attributes..."

“in my gigabytes of data?”

Now we can help you!

For the first question:

“What things should I anonymize?”

We wrote this paper:

“Protecting Privacy in Software Logs: What should be Anonymized?”

This paper got accepted in FSE 2025.

And concluded these attributes as generally sensitive information:

IP address

MAC address

Host name

File path

ID

URL

Username

Port number

Configuration details

Let’s see some examples.

Attribute	Example
IP Address	*Invalid user webmaster from 173.234.31.186*
MAC address	*ARPT: 621131.293163: wl0: Roamed or switched channel, reason #8, bssid 5c:50:15:4c:18:13, last RSSI -64*
Host name	*proxy.cse.cuhk.edu.hk: 5070 close, 0 bytes sent, 0 bytes received, lifetime 00:01*
File path	*workerEnv.init() ok /etc/httpd/conf/workers2.properties*
ID	*Verification succeeded for blk_-4980916519894289629*
URL	*the url = http://baike.baidu.com/item/%E8%93%9D%E9%87%87%E5%92%8C/462624?fr=aladdin*
Username	*Invalid user webmaster from 173.234.31.186*
Port number	*proxy.cse.cuhk.edu.hk: 5070 close, 0 bytes sent, 0 bytes received, lifetime 00:01*
Configuration details	*mapResourceRequest:<memory:1024, vCores:1>*

Now that you have the answer to your first question, Let’s move forward to the second one:

“How should I find all these attributes in my gigabytes of data?”

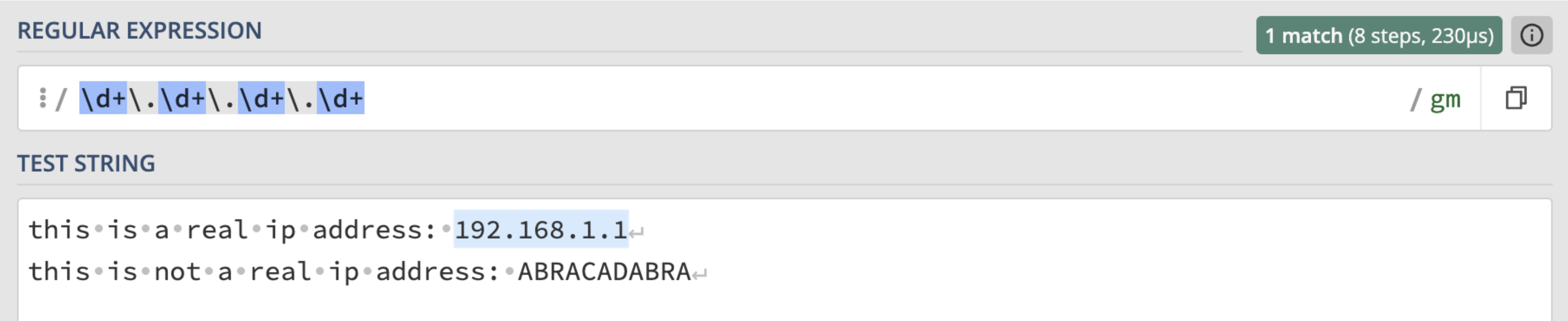
Maybe
regular expressions?

Let’s say we have this IP address:

192.168.1.1

Then we use this regular expression:

```
\d+\.\d+\.\d+\.\d+
```



REGULAR EXPRESSION

1 match (8 steps, 230µs)

```
/ \d+\.\d+\.\d+\.\d+ / gm
```

TEST STRING

```
this is a real ip address: 192.168.1.1
this is not a real ip address: ABRACADABRA
```

Voilà!

Right?

Well…

Not that much.

	IP address	**P (%)**	**R (%)**	**F1 (%)**	**Source**
1	`(/|)([0-9]+\.){3}[0-9]+(:[0-9]+|)(:|)`	**92.1**	85.1	88.4	[21, 35, 104, 106]
2	`([0-9.]+)\s`	11.4	48.8	18.5	[110]
3	`([0-9]+.){3}[0-9]+(:[0-9]+)`	23.0	0.6	1.1	[37]
4	`((\d+).(\d+).(\d+).(\d+))`	32.5	**99.7**	49.0	[107]
5	`(\d)+3\d(:\d+)?`	8.6	9.6	9.1	[72]
6	`(\d+\.){3}\d+(:\d+)?`	**92.1**	85.1	88.4	[65]
7	`^(25[0-5]|2[0-4]\d|[0-1]?\d?\d)(\.(25[0-5]|2[0-4]\d|[0-1]?\d?\d)){3}$`	0.0	0.0	0.0	[71]
8	`(\b\d{1,3}(?:\.\d{1,3}){3}\b)`	**92.1**	85.1	**88.5**	[98]
9	`(\d{1,3}(?:\.\d{1,3}){3}):?\d*`	**92.1**	85.1	**88.5**	[98]
10	`\d+\.\d+\.\d+\.\d+`	**92.1**	85.1	88.4	[54]
11	`(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})[,: )]`	90.7	78.6	84.2	[55]
12	`[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}`	31.5	**99.7**	47.9	[22]
13	`(/|)(\d+.){3}\d+(:\d+)?`	32.5	**99.7**	49.1	[83]
14	`[0-9]+\.[0-9\.:]*[0-9]`	56.7	85.1	68.1	[56]
15	`(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})`	**92.1**	85.1	**88.5**	Company 1
16	`\b\d{1,3}(?:\.\d{1,3}){2,}\b`	91.8	85.1	88.3	Company 2
17	`(\b\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\b)`	**92.1**	85.1	**88.5**	Company 3

There is NO common ground truth for regular expressions!

	IP address	P (%)	R (%)	F1 (%)	Source
1	(/|)([0-9]+\.){3}[0-9]+(:[0-9]+|)(:|)	**92.1**	85.1	88.4	[21, 35, 104, 106]
2	([0-9.]+)\s	11.4	48.8	18.5	[110]
3	([0-9]+.){3}[0-9]+(:[0-9]+)	23.0	0.6	1.1	[37]
4	((\d+).(\d+).(\d+).(\d+))	32.5	**99.7**	49.0	[107]
5	(\d)+3\d(:\d+)?	8.6	9.6	9.1	[72]
6	(\d+\.){3}\d+(:\d+)?	**92.1**	85.1	88.4	[65]
7	^(25[0-5]|2[0-4]\d|[0-1]?\d?\d)(\.(25[0-5]|2[0-4]\d|[0-1]?\d?\d)){3}$	0.0	0.0	0.0	[71]
8	(\b\d{1,3}(?:\.\d{1,3}){3}\b)	**92.1**	85.1	**88.5**	[98]
9	(\d{1,3}(?:\.\d{1,3}){3}):?\d*	**92.1**	85.1	**88.5**	[98]
10	\d+\.\d+\.\d+\.\d+	**92.1**	85.1	88.4	[54]
11	(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})[,:)]	90.7	78.6	84.2	[55]
12	[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}	31.5	**99.7**	47.9	[22]
13	(/|)(\d+.){3}\d+(:\d+)?	32.5	**99.7**	49.1	[83]
14	[0-9]+\.[0-9\.:]*[0-9]	56.7	85.1	68.1	[56]
15	(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})	**92.1**	85.1	**88.5**	Company 1
16	\b\d{1,3}(?:\.\d{1,3}){2,}\b	91.8	85.1	88.3	Company 2
17	(\b\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\b)	**92.1**	85.1	**88.5**	Company 3

<table>
<tr><th></th><th>IP address</th><th>P (%)</th><th>R (%)</th><th>F1 (%)</th><th>Source</th></tr>
<tr><td>1</td><td>(/|)([0-9]+\.){3}[0-9]+(:[0-9]+|)(:|)</td><td><b>92.1</b></td><td>85.1</td><td>88.4</td><td>[21, 35, 104, 106]</td></tr>
<tr><td>2</td><td>([0-9.]+)\s</td><td>11.4</td><td>48.8</td><td>18.5</td><td>[110]</td></tr>
<tr><td>3</td><td>([0-9]+.){3}[0-9]+(:[0-9]+)</td><td>23.0</td><td>0.6</td><td>1.1</td><td>[37]</td></tr>
<tr><td>4</td><td>((\d+).(\d+).(\d+).(\d+))</td><td>32.5</td><td><b>99.7</b></td><td>49.0</td><td>[107]</td></tr>
<tr><td>5</td><td>(\d)+3\d(:\d+)?</td><td>8.6</td><td>9.6</td><td>9.1</td><td>[72]</td></tr>
<tr><td>6</td><td>(\d+\.){3}\d+(:\d+)?</td><td><b>92.1</b></td><td>85.1</td><td>88.4</td><td>[65]</td></tr>
<tr><td>7</td><td>^(25[0-5]|2[0-4]\d|[0-1]?\d?\d)(\.(25[0-5]|2[0-4]\d|[0-1]?\d?\d)){3}$</td><td>0.0</td><td>0.0</td><td>0.0</td><td>[71]</td></tr>
<tr><td>8</td><td>(\b\d{1,3}(?:\.\d{1,3}){3}\b)</td><td><b>92.1</b></td><td>85.1</td><td><b>88.5</b></td><td>[98]</td></tr>
<tr><td>9</td><td>(\d{1,3}(?:\.\d{1,3}){3}):?\d*</td><td><b>92.1</b></td><td>85.1</td><td><b>88.5</b></td><td>[98]</td></tr>
<tr><td>10</td><td>\d+\.\d+\.\d+\.\d+</td><td><b>92.1</b></td><td>85.1</td><td>88.4</td><td>[54]</td></tr>
<tr><td>11</td><td>(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})[,:)]</td><td>90.7</td><td>78.6</td><td>84.2</td><td>[55]</td></tr>
<tr><td>12</td><td>[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}</td><td>31.5</td><td><b>99.7</b></td><td>47.9</td><td>[22]</td></tr>
<tr><td>13</td><td>(/|)(\d+.){3}\d+(:\d+)?</td><td>32.5</td><td><b>99.7</b></td><td>49.1</td><td>[83]</td></tr>
<tr><td>14</td><td>[0-9]+\.[0-9\.:]*[0-9]</td><td>56.7</td><td>85.1</td><td>68.1</td><td>[56]</td></tr>
<tr><td>15</td><td>(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})</td><td><b>92.1</b></td><td>85.1</td><td><b>88.5</b></td><td>Company 1</td></tr>
<tr><td>16</td><td>\b\d{1,3}(?:\.\d{1,3}){2,}\b</td><td>91.8</td><td>85.1</td><td>88.3</td><td>Company 2</td></tr>
<tr><td>17</td><td>(\b\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\b)</td><td><b>92.1</b></td><td>85.1</td><td><b>88.5</b></td><td>Company 3</td></tr>
</table>

<table>
<tr><th></th><th>IP address</th><th>P (%)</th><th>R (%)</th><th>F1 (%)</th><th>Source</th></tr>
<tr><td>1</td><td>(/|)([0-9]+\.){3}[0-9]+(:[0-9]+|)(:|)</td><td><b>92.1</b></td><td>85.1</td><td>88.4</td><td>[21, 35, 104, 106]</td></tr>
<tr><td>2</td><td>([0-9.]+)\s</td><td>11.4</td><td>48.8</td><td>18.5</td><td>[110]</td></tr>
<tr><td>3</td><td>([0-9]+.){3}[0-9]+(:[0-9]+)</td><td>23.0</td><td>0.6</td><td>1.1</td><td>[37]</td></tr>
<tr><td>4</td><td>((\d+).(\d+).(\d+).(\d+))</td><td>32.5</td><td><b>99.7</b></td><td>49.0</td><td>[107]</td></tr>
<tr><td>5</td><td>(\d)+3\d(:\d+)?</td><td>8.6</td><td>9.6</td><td>9.1</td><td>[72]</td></tr>
<tr><td>6</td><td>(\d+\.){3}\d+(:\d+)?</td><td><b>92.1</b></td><td>85.1</td><td>88.4</td><td>[65]</td></tr>
<tr><td>7</td><td>^(25[0-5]|2[0-4]\d|[0-1]?\d?\d)(\.(25[0-5]|2[0-4]\d|[0-1]?\d?\d)){3}$</td><td>0.0</td><td>0.0</td><td>0.0</td><td>[71]</td></tr>
<tr><td>8</td><td>(\b\d{1,3}(?:\.\d{1,3}){3}\b)</td><td><b>92.1</b></td><td>85.1</td><td><b>88.5</b></td><td>[98]</td></tr>
<tr><td>9</td><td>(\d{1,3}(?:\.\d{1,3}){3}):?\d*</td><td><b>92.1</b></td><td>85.1</td><td><b>88.5</b></td><td>[98]</td></tr>
<tr><td>10</td><td>\d+\.\d+\.\d+\.\d+</td><td><b>92.1</b></td><td>85.1</td><td>88.4</td><td>[54]</td></tr>
<tr><td>11</td><td>(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})[,:)]</td><td>90.7</td><td>78.6</td><td>84.2</td><td>[55]</td></tr>
<tr><td>12</td><td>[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}</td><td>31.5</td><td><b>99.7</b></td><td>47.9</td><td>[22]</td></tr>
<tr><td>13</td><td>(/|)(\d+.){3}\d+(:\d+)?</td><td>32.5</td><td><b>99.7</b></td><td>49.1</td><td>[83]</td></tr>
<tr><td>14</td><td>[0-9]+\.[0-9\.:]*[0-9]</td><td>56.7</td><td>85.1</td><td>68.1</td><td>[56]</td></tr>
<tr><td>15</td><td>(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})</td><td><b>92.1</b></td><td>85.1</td><td><b>88.5</b></td><td>Company 1</td></tr>
<tr><td>16</td><td>\b\d{1,3}(?:\.\d{1,3}){2,}\b</td><td>91.8</td><td>85.1</td><td>88.3</td><td>Company 2</td></tr>
<tr><td>17</td><td>(\b\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\b)</td><td><b>92.1</b></td><td>85.1</td><td><b>88.5</b></td><td>Company 3</td></tr>
</table>

<table>
<thead>
<tr><th></th><th>IP address</th><th>P (%)</th><th>R (%)</th><th>F1 (%)</th><th>Source</th></tr>
</thead>
<tbody>
<tr><td>1</td><td>(/|)([0-9]+\.){3}[0-9]+(:[0-9]+|)(:|)</td><td><b>92.1</b></td><td>85.1</td><td>88.4</td><td>[21, 35, 104, 106]</td></tr>
<tr><td>2</td><td>([0-9.]+)\s</td><td>11.4</td><td>48.8</td><td>18.5</td><td>[110]</td></tr>
<tr><td>3</td><td>([0-9]+.){3}[0-9]+(:[0-9]+)</td><td>23.0</td><td>0.6</td><td>1.1</td><td>[37]</td></tr>
<tr><td>4</td><td>((\d+).(\d+).(\d+).(\d+))</td><td>32.5</td><td><b>99.7</b></td><td>49.0</td><td>[107]</td></tr>
<tr><td>5</td><td>(\d)+3\d(:\d+)?</td><td>8.6</td><td>9.6</td><td>9.1</td><td>[72]</td></tr>
<tr><td>6</td><td>(\d+\.){3}\d+(:\d+)?</td><td><b>92.1</b></td><td>85.1</td><td>88.4</td><td>[65]</td></tr>
<tr><td>7</td><td>^(25[0-5]|2[0-4]\d|[0-1]?\d?\d)(\.(25[0-5]|2[0-4]\d|[0-1]?\d?\d)){3}$</td><td>0.0</td><td>0.0</td><td>0.0</td><td>[71]</td></tr>
<tr><td>8</td><td>(\b\d{1,3}(?:\.\d{1,3}){3}\b)</td><td><b>92.1</b></td><td>85.1</td><td><b>88.5</b></td><td>[98]</td></tr>
<tr><td>9</td><td>(\d{1,3}(?:\.\d{1,3}){3}):?\d*</td><td><b>92.1</b></td><td>85.1</td><td><b>88.5</b></td><td>[98]</td></tr>
<tr><td>10</td><td>\d+\.\d+\.\d+\.\d+</td><td><b>92.1</b></td><td>85.1</td><td>88.4</td><td>[54]</td></tr>
<tr><td>11</td><td>(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})[,:)]</td><td>90.7</td><td>78.6</td><td>84.2</td><td>[55]</td></tr>
<tr><td>12</td><td>[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}</td><td>31.5</td><td><b>99.7</b></td><td>47.9</td><td>[22]</td></tr>
<tr><td>13</td><td>(/|)(\d+.){3}\d+(:\d+)?</td><td>32.5</td><td><b>99.7</b></td><td>49.1</td><td>[83]</td></tr>
<tr><td>14</td><td>[0-9]+\.[0-9\.:]*[0-9]</td><td>56.7</td><td>85.1</td><td>68.1</td><td>[56]</td></tr>
<tr><td>15</td><td>(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})</td><td><b>92.1</b></td><td>85.1</td><td><b>88.5</b></td><td>Company 1</td></tr>
<tr><td>16</td><td>\b\d{1,3}(?:\.\d{1,3}){2,}\b</td><td>91.8</td><td>85.1</td><td>88.3</td><td>Company 2</td></tr>
<tr><td>17</td><td>(\b\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\b)</td><td><b>92.1</b></td><td>85.1</td><td><b>88.5</b></td><td>Company 3</td></tr>
</tbody>
</table>

<table>
<tr><th></th><th>IP address</th><th>P (%)</th><th>R (%)</th><th>F1 (%)</th><th>Source</th></tr>
<tr><td>1</td><td>(/|)([0-9]+\.){3}[0-9]+(:[0-9]+|)(:|)</td><td>**92.1**</td><td>85.1</td><td>88.4</td><td>[21, 35, 104, 106]</td></tr>
<tr><td>2</td><td>([0-9.]+)\s</td><td>11.4</td><td>48.8</td><td>18.5</td><td>[110]</td></tr>
<tr><td>3</td><td>([0-9]+.){3}[0-9]+(:[0-9]+)</td><td>23.0</td><td>0.6</td><td>1.1</td><td>[37]</td></tr>
<tr><td>4</td><td>((\d+).(\d+).(\d+).(\d+))</td><td>32.5</td><td>**99.7**</td><td>49.0</td><td>[107]</td></tr>
<tr><td>5</td><td>(\d)+3\d(:\d+)?</td><td>8.6</td><td>9.6</td><td>9.1</td><td>[72]</td></tr>
<tr><td>6</td><td>(\d+\.){3}\d+(:\d+)?</td><td>**92.1**</td><td>85.1</td><td>88.4</td><td>[65]</td></tr>
<tr><td>7</td><td>^(25[0-5]|2[0-4]\d|[0-1]?\d?\d)(\.(25[0-5]|2[0-4]\d|[0-1]?\d?\d)){3}$</td><td>0.0</td><td>0.0</td><td>0.0</td><td>[71]</td></tr>
<tr><td>8</td><td>(\b\d{1,3}(?:\.\d{1,3}){3}\b)</td><td>**92.1**</td><td>85.1</td><td>**88.5**</td><td>[98]</td></tr>
<tr><td>9</td><td>(\d{1,3}(?:\.\d{1,3}){3}):?\d*</td><td>**92.1**</td><td>85.1</td><td>**88.5**</td><td>[98]</td></tr>
<tr><td>10</td><td>\d+\.\d+\.\d+\.\d+</td><td>**92.1**</td><td>85.1</td><td>88.4</td><td>[54]</td></tr>
<tr><td>11</td><td>(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})[,:)]</td><td>90.7</td><td>78.6</td><td>84.2</td><td>[55]</td></tr>
<tr><td>12</td><td>[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}</td><td>31.5</td><td>**99.7**</td><td>47.9</td><td>[22]</td></tr>
<tr><td>13</td><td>(/|)(\d+.){3}\d+(:\d+)?</td><td>32.5</td><td>**99.7**</td><td>49.1</td><td>[83]</td></tr>
<tr><td>14</td><td>[0-9]+\.[0-9\.:]*[0-9]</td><td>56.7</td><td>85.1</td><td>68.1</td><td>[56]</td></tr>
<tr><td>15</td><td>(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})</td><td>**92.1**</td><td>85.1</td><td>**88.5**</td><td>Company 1</td></tr>
<tr><td>16</td><td>\b\d{1,3}(?:\.\d{1,3}){2,}\b</td><td>91.8</td><td>85.1</td><td>88.3</td><td>Company 2</td></tr>
<tr><td>17</td><td>(\b\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\b)</td><td>**92.1**</td><td>85.1</td><td>**88.5**</td><td>Company 3</td></tr>
</table>

	IP address	P (%)	R (%)	F1 (%)	Source
1	`(/|)([0-9]+\.){3}[0-9]+(:[0-9]+|)(:|)`	**92.1**	85.1	88.4	[21, 35, 104, 106]
2	`([0-9.]+)\s`	11.4	48.8	18.5	[110]
3	`([0-9]+.){3}[0-9]+(:[0-9]+)`	23.0	0.6	1.1	[37]
4	`((\d+).(\d+).(\d+).(\d+))`	32.5	**99.7**	49.0	[107]
5	`(\d)+3\d(:\d+)?`	8.6	9.6	9.1	[72]
6	`(\d+\.){3}\d+(:\d+)?`	**92.1**	85.1	88.4	[65]
7	`^(25[0-5]|2[0-4]\d|[0-1]?\d?\d)(\.(25[0-5]|2[0-4]\d|[0-1]?\d?\d)){3}$`	0.0	0.0	0.0	[71]
8	`(\b\d{1,3}(?:\.\d{1,3}){3}\b)`	**92.1**	85.1	**88.5**	[98]
9	`(\d{1,3}(?:\.\d{1,3}){3}):?\d*`	**92.1**	85.1	**88.5**	[98]
10	`\d+\.\d+\.\d+\.\d+`	**92.1**	85.1	88.4	[54]
11	`(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})[,: )]`	90.7	78.6	84.2	[55]
12	`[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}`	31.5	**99.7**	47.9	[22]
13	`(/|)(\d+.){3}\d+(:\d+)?`	32.5	**99.7**	49.1	[83]
14	`[0-9]+\.[0-9\.:]*[0-9]`	56.7	85.1	68.1	[56]
15	`(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})`	**92.1**	85.1	**88.5**	Company 1
16	`\b\d{1,3}(?:\.\d{1,3}){2,}\b`	91.8	85.1	88.3	Company 2
17	`(\b\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\b)`	**92.1**	85.1	**88.5**	Company 3

	IP address	**P (%)**	**R (%)**	**F1 (%)**	**Source**
1	`(/|)([0-9]+\.){3}[0-9]+(:[0-9]+|)(:|)`	**92.1**	85.1	88.4	[21, 35, 104, 106]
2	`([0-9.]+)\s`	11.4	48.8	18.5	[110]
3	`([0-9]+.){3}[0-9]+(:[0-9]+)`	23.0	0.6	1.1	[37]
4	`((\d+).(\d+).(\d+).(\d+))`	32.5	**99.7**	49.0	[107]
5	`(\d)+3\d(:\d+)?`	8.6	9.6	9.1	[72]
6	`(\d+\.){3}\d+(:\d+)?`	**92.1**	85.1	88.4	[65]
7	`^(25[0-5]|2[0-4]\d|[0-1]?\d?\d)(\.(25[0-5]|2[0-4]\d|[0-1]?\d?\d)){3}$`	0.0	0.0	0.0	[71]
8	`(\b\d{1,3}(?:\.\d{1,3}){3}\b)`	**92.1**	85.1	**88.5**	[98]
9	`(\d{1,3}(?:\.\d{1,3}){3}):?\d*`	**92.1**	85.1	**88.5**	[98]
10	`\d+\.\d+\.\d+\.\d+`	**92.1**	85.1	88.4	[54]
11	`(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})[,: )]`	90.7	78.6	84.2	[55]
12	`[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}.[0-9]{1,3}`	31.5	**99.7**	47.9	[22]
13	`(/|)(\d+.){3}\d+(:\d+)?`	32.5	**99.7**	49.1	[83]
14	`[0-9]+\.[0-9\.:]*[0-9]`	56.7	85.1	68.1	[56]
15	`(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})`	**92.1**	85.1	**88.5**	Company 1
16	`\b\d{1,3}(?:\.\d{1,3}){2,}\b`	91.8	85.1	88.3	Company 2
17	`(\b\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\.)(\d{1,3}\b)`	**92.1**	85.1	**88.5**	Company 3

Small differences in regex design have a LARGE impact!

<table>
<tr><th></th><th>File path</th><th>P (%)</th><th>R (%)</th><th>F1 (%)</th><th>Source</th></tr>
<tr><td>1</td><td>((((?<!\w)[A-Z,a-z]:)|(\.{1,2}\\))([^\b%\/\|:\n\"]*))|(\"\2([^%\/\|:\n\"]*)\")|((?<!\w)(\.{1,2})?(?<!\/)(\/((\\\b)|[^\b%\|:\n\"\\\/])+)+\/?)</td><td>59.0</td><td>98.3</td><td>73.7</td><td>[55]</td></tr>
<tr><td>2</td><td>/[\w/. :-]+</td><td>55.6</td><td>99.5</td><td>71.3</td><td>[97]</td></tr>
<tr><td>3</td><td>(/[^/\s]+)+</td><td>48.1</td><td>99.5</td><td>64.9</td><td>[97]</td></tr>
<tr><td>4</td><td>(([A-Z]:)|)(/\S+)+</td><td>47.8</td><td>99.5</td><td>64.5</td><td>[104]</td></tr>
<tr><td>5</td><td>(/|)(([\w.-]+|\<*\>)/)+([\w.-]+|\<*\>)</td><td><b>66.7</b></td><td>98.1</td><td><b>79.4</b></td><td>[83]</td></tr>
<tr><td>6</td><td>([A-Za-z]:|\.){0,1}(/|\\)[0-9A-Za-z\-_\.:/*\+\$#@!\\\?=%&]+(?<![:\.])</td><td>48.1</td><td><b>100.0</b></td><td>65.0</td><td>[56]</td></tr>
<tr><td>7</td><td>\/(\S+)</td><td>47.8</td><td>99.5</td><td>64.5</td><td>Company 3</td></tr>
</table>

<table>
<tr><th></th><th>URL</th><th>P (%)</th><th>R (%)</th><th>F1 (%)</th><th>Source</th></tr>
<tr><td>1</td><td>[A-Za-z\.]+://[A-Za-z0-9\.\/\+#@:_\-]+(?<![:\.])</td><td>91.4</td><td>99.2</td><td>**95.1**</td><td>[56]</td></tr>
<tr><td>2</td><td>(https?://\S+)</td><td>**100.0**</td><td>39.1</td><td>56.2</td><td>[55]</td></tr>
<tr><td>3</td><td>https?://[^\s#]+#[A-Za-z0-9\-\=\+]+</td><td>0.0</td><td>0.0</td><td>0.0</td><td>[97]</td></tr>
<tr><td>4</td><td>http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\\(\\).]|(?:\%[0-9afA-F][0-9a-fA-F]))+</td><td>**100.0**</td><td>39.1</td><td>56.2</td><td>[97]</td></tr>
<tr><td>5</td><td>([\w-]+\.)+[\w-]+(:\d+)?</td><td>0.9</td><td>**100.0**</td><td>1.8</td><td>[108]</td></tr>
<tr><td>6</td><td>(\S+\.\S+(\.\S+)+(:\d+)?)|(\w+-\w+(-\w+)+)</td><td>0.7</td><td>**100.0**</td><td>1.4</td><td>[104]</td></tr>
<tr><td>7</td><td>\bhttps?://(www.)?[a-zA-Z0-9-]+(.[a-zA-Z]{2,})+(:[0-9]{1,5})?(/[^\s]*)?\b</td><td>**100.0**</td><td>31.2</td><td>47.6</td><td>[83]</td></tr>
</table>

	ID	P (%)	R (%)	F1 (%)	Source
1	`(?:UUID|GUID|version|id)[\\=:\"\’\s]*\b[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}\b`	0.0	0.0	0.0	[97]
2	`<([^>]+)>`	2.6	0.1	0.2	[97]
3	`[pP]id[:|-|=|\s/]*(\d+)`	97.7	1.3	2.5	[55]
4	`[uU]id[:|-|=|\s/]*(\d+)`	**99.8**	**23.5**	**38.1**	[55]

<table>
<tr><th></th><th>ID</th><th>P (%)</th><th>R (%)</th><th>F1 (%)</th><th>Source</th></tr>
<tr><td>1</td><td>(?:UUID|GUID|version|id)[\\=:\"\'\s]*\b[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}\b</td><td>0.0</td><td>0.0</td><td>0.0</td><td>[97]</td></tr>
<tr><td>2</td><td><([^>]+)></td><td>2.6</td><td>0.1</td><td>0.2</td><td>[97]</td></tr>
<tr><td>3</td><td>[pP]id[:|-|=|\s/]*(\d+)</td><td>97.7</td><td>1.3</td><td>2.5</td><td>[55]</td></tr>
<tr><td>4</td><td>[uU]id[:|-|=|\s/]*(\d+)</td><td>**99.8**</td><td>**23.5**</td><td>**38.1**</td><td>[55]</td></tr>
</table>

<table>
<tr><th></th><th>Username</th><th>P (%)</th><th>R (%)</th><th>F1 (%)</th><th>Source</th></tr>
<tr><td>1</td><td>user(|)[A-Za-z0-9]+(?<!request)(?! methods)</td><td>**42.0**</td><td>25.3</td><td>31.6</td><td>[56]</td></tr>
<tr><td>2</td><td>user\:\s(\w+)</td><td>0.0</td><td>0.0</td><td>0.0</td><td>[105]</td></tr>
<tr><td>3</td><td>r?[uU]ser[:|-|=|\s/]*<(\w+)>|r?[uU]ser[:|-|=|\s/]*(\w+)</td><td>35.2</td><td>**72.0**</td><td>**47.3**</td><td>[55]</td></tr>
</table>

<table>
<tr><th></th><th>ID</th><th>P (%)</th><th>R (%)</th><th>F1 (%)</th><th>Source</th></tr>
<tr><td>1</td><td>(?:UUID|GUID|version|id)[\\=:\"\'\s]*\b[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}\b</td><td>0.0</td><td>0.0</td><td>0.0</td><td>[97]</td></tr>
<tr><td>2</td><td>&lt;([^&gt;]+)&gt;</td><td>2.6</td><td>0.1</td><td>0.2</td><td>[97]</td></tr>
<tr><td>3</td><td>[pP]id[:|-|=|\s/]*(\d+)</td><td>97.7</td><td>1.3</td><td>2.5</td><td>[55]</td></tr>
<tr><td>4</td><td>[uU]id[:|-|=|\s/]*(\d+)</td><td>**99.8**</td><td>**23.5**</td><td>**38.1**</td><td>[55]</td></tr>
</table>

<table>
<tr><th></th><th>Username</th><th>P (%)</th><th>R (%)</th><th>F1 (%)</th><th>Source</th></tr>
<tr><td>1</td><td>user(|)[A-Za-z0-9]+(?&lt;!request)(?! methods)</td><td>**42.0**</td><td>25.3</td><td>31.6</td><td>[56]</td></tr>
<tr><td>2</td><td>user\:\s(\w+)</td><td>0.0</td><td>0.0</td><td>0.0</td><td>[105]</td></tr>
<tr><td>3</td><td>r?[uU]ser[:|-|=|\s/]*&lt;(\w+)&gt;|r?[uU]ser[:|-|=|\s/]*(\w+)</td><td>35.2</td><td>**72.0**</td><td>**47.3**</td><td>[55]</td></tr>
</table>

<table>
<tr><th></th><th>Port</th><th>P (%)</th><th>R (%)</th><th>F1 (%)</th><th>Source</th></tr>
<tr><td>1</td><td>[pP]ort[=: |:|=|: |\s/]*(\d1,5)</td><td>**96.0**</td><td>**8.1**</td><td>**15.0**</td><td>[55]</td></tr>
</table>

	ID	**P (%)**	**R (%)**	**F1 (%)**	**Source**
1	`(?:UUID|GUID|version|id)[\\=:\"\'\s]*\b[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}\b`	0.0	0.0	0.0	[97]
2	`<([^>]+)>`	2.6	0.1	0.2	[97]
3	`[pP]id[:|-|=|\s/]*(\d+)`	97.7	1.3	2.5	[55]
4	`[uU]id[:|-|=|\s/]*(\d+)`	**99.8**	**23.5**	**38.1**	[55]

	Username	**P (%)**	**R (%)**	**F1 (%)**	**Source**
1	`user( | )[A-Za-z0-9]+(?<!request)(?! methods)`	**42.0**	25.3	31.6	[56]
2	`user\:\s(\w+)`	0.0	0.0	0.0	[105]
3	`r?[uU]ser[:|-|=|\s/]*<(\w+)>|r?[uU]ser[:|-|=|\s/]*(\w+)`	35.2	**72.0**	**47.3**	[55]

	Port	**P (%)**	**R (%)**	**F1 (%)**	**Source**
1	`[pP]ort[=: |:|=|: |\s/]*(\d1,5)`	**96.0**	**8.1**	**15.0**	[55]

	Configuration details	**P (%)**	**R (%)**	**F1 (%)**	**Source**
1	`size\s+(\d+)`	**19.2**	**14.2**	**16.3**	[98]

By the way,
did you know that the
order of regexes
changes the results, …

A LOT?

<table>
<tr><th rowspan="2"><b><i>Attribute</i></b></th><th rowspan="2"><b><i>Best Regex Pattern</i></b></th><th colspan="2"><b><i>Percision (%)</i></b></th><th colspan="2"><b><i>Recall (%)</i></b></th><th colspan="2"><b><i>F1 (%)</i></b></th></tr>
<tr><th><b><i>Min</i></b></th><th><b><i>Max</i></b></th><th><b><i>Min</i></b></th><th><b><i>Max</i></b></th><th><b><i>Min</i></b></th><th><b><i>Max</i></b></th></tr>
<tr><td><i>IP</i></td><td>(\b\d{1,3}(?:\.\d{1,3}){3}\b)</td><td><i>92.1</i></td><td><i>93.0</i></td><td><i>85.1</i></td><td><i>85.1</i></td><td><i>88.5</i></td><td><i>88.9</i></td></tr>
<tr><td><i>MAC</i></td><td>\([0-9A-Fa-f]{2}[:-]){5}([0-9A-Fa-f]{2})\</td><td><i>98.6</i></td><td><i>98.6</i></td><td><i>100.0</i></td><td><i>100.0</i></td><td><i>99.3</i></td><td><i>99.3</i></td></tr>
<tr><td><i>File path</i></td><td>(/|)(([\w.-]+|\<*\>)/)+([\w.-]+|\<*\>)</td><td><i>66.6</i></td><td><i>68.4</i></td><td><i>97.7</i></td><td><i>98.0</i></td><td><i>79.3</i></td><td><i>80.5</i></td></tr>
<tr><td><i>ID</i></td><td>[uU]id[:|-|=|\s/]*(\d+)</td><td><i>99.8</i></td><td><i>99.8</i></td><td><i>23.5</i></td><td><i>23.5</i></td><td><i>38.0</i></td><td><i>38.0</i></td></tr>
<tr><td><i>URL</i></td><td>[A-Za-z\.]+://[A-Za-z0-9\.\/\+#@:_\-]+(?<![:\.])</td><td><i>70.6</i></td><td><i>94.6</i></td><td><i>9.4</i></td><td><i>99.2</i></td><td><i>16.6</i></td><td><i>95.1</i></td></tr>
<tr><td><i>Username</i></td><td>r?[uU]ser[:|-|=|\s/]*<(\w+)>|r?[uU]ser[:|-|=|\s/]*(\w+)</td><td><i>36.6</i></td><td><i>36.6</i></td><td><i>72.0</i></td><td><i>72.0</i></td><td><i>48.5</i></td><td><i>48.5</i></td></tr>
<tr><td><i>Port</i></td><td>[pP]ort[=: |:|=|: |\s/]*(\d1,5)</td><td><i>96.0</i></td><td><i>96.2</i></td><td><i>8.1</i></td><td><i>8.1</i></td><td><i>15.0</i></td><td><i>15.0</i></td></tr>
<tr><td><i>Configuration</i></td><td>size\s+(\d+)</td><td><i>19.2</i></td><td><i>19.2</i></td><td><i>14.2</i></td><td><i>14.2</i></td><td><i>16.3</i></td><td><i>16.3</i></td></tr>
</table>

<table>
<tr><th rowspan="2">Attribute</th><th rowspan="2">Best Regex Pattern</th><th colspan="2">Percision (%)</th><th colspan="2">Recall (%)</th><th colspan="2">F1 (%)</th></tr>
<tr><th>Min</th><th>Max</th><th>Min</th><th>Max</th><th>Min</th><th>Max</th></tr>
<tr><td><em>IP</em></td><td>(\b\d{1,3}(?:\.\d{1,3}){3}\b)</td><td><em>92.1</em></td><td><em>93.0</em></td><td><em>85.1</em></td><td><em>85.1</em></td><td><em>88.5</em></td><td><em>88.9</em></td></tr>
<tr><td><em>MAC</em></td><td>\([0-9A-Fa-f]{2}[:-]){5}([0-9A-Fa-f]{2})\</td><td><em>98.6</em></td><td><em>98.6</em></td><td><em>100.0</em></td><td><em>100.0</em></td><td><em>99.3</em></td><td><em>99.3</em></td></tr>
<tr><td><em>File path</em></td><td>(/|)(([\w.-]+|\<*\>)/)+([\w.-]+|\<*\>)</td><td><em>66.6</em></td><td><em>68.4</em></td><td><em>97.7</em></td><td><em>98.0</em></td><td><em>79.3</em></td><td><em>80.5</em></td></tr>
<tr><td><em>ID</em></td><td>[uU]id[:|-|=|\s/]*(\d+)</td><td><em>99.8</em></td><td><em>99.8</em></td><td><em>23.5</em></td><td><em>23.5</em></td><td><em>38.0</em></td><td><em>38.0</em></td></tr>
<tr><td><em>URL</em></td><td>[A-Za-z\.]+://[A-Za-z0-9\.\/\+#@:_\-]+(?<![:\.])</td><td><em>70.6</em></td><td><em>94.6</em></td><td><em>9.4</em></td><td><em>99.2</em></td><td><em>16.6</em></td><td><em>95.1</em></td></tr>
<tr><td><em>Username</em></td><td>r?[uU]ser[:|-|=|\s/]*<(\w+)>|r?[uU]ser[:|-|=|\s/]*(\w+)</td><td><em>36.6</em></td><td><em>36.6</em></td><td><em>72.0</em></td><td><em>72.0</em></td><td><em>48.5</em></td><td><em>48.5</em></td></tr>
<tr><td><em>Port</em></td><td>[pP]ort[=: |:|=|: |\s/]*(\d1,5)</td><td><em>96.0</em></td><td><em>96.2</em></td><td><em>8.1</em></td><td><em>8.1</em></td><td><em>15.0</em></td><td><em>15.0</em></td></tr>
<tr><td><em>Configuration</em></td><td>size\s+(\d+)</td><td><em>19.2</em></td><td><em>19.2</em></td><td><em>14.2</em></td><td><em>14.2</em></td><td><em>16.3</em></td><td><em>16.3</em></td></tr>
</table>

<table>
<tr><th></th><th></th><th colspan="2">Percision (%)</th><th colspan="2">Recall (%)</th><th colspan="2">F1 (%)</th></tr>
<tr><th>Attribute</th><th>Best Regex Pattern</th><th>Min</th><th>Max</th><th>Min</th><th>Max</th><th>Min</th><th>Max</th></tr>
<tr><td>IP</td><td>(\b\d{1,3}(?:\.\d{1,3}){3}\b)</td><td>92.1</td><td>93.0</td><td>85.1</td><td>85.1</td><td>88.5</td><td>88.9</td></tr>
<tr><td>MAC</td><td>\([0-9A-Fa-f]{2}[:-]){5}([0-9A-Fa-f]{2})\</td><td>98.6</td><td>98.6</td><td>100.0</td><td>100.0</td><td>99.3</td><td>99.3</td></tr>
<tr><td>File path</td><td>(/|)(([\w.-]+|\<*\>)/)+([\w.-]+|\<*\>)</td><td>66.6</td><td>68.4</td><td>97.7</td><td>98.0</td><td>79.3</td><td>80.5</td></tr>
<tr><td>ID</td><td>[uU]id[:|-|=|\s/]*(\d+)</td><td>99.8</td><td>99.8</td><td>23.5</td><td>23.5</td><td>38.0</td><td>38.0</td></tr>
<tr><td>URL</td><td>[A-Za-z\.]+://[A-Za-z0-9\.\/\+#@:_\-]+(?<![:\.])</td><td>70.6</td><td>94.6</td><td>9.4</td><td>99.2</td><td>16.6</td><td>95.1</td></tr>
<tr><td>Username</td><td>r?[uU]ser[:|-|=|\s/]*<(\w+)>|r?[uU]ser[:|-|=|\s/]*(\w+)</td><td>36.6</td><td>36.6</td><td>72.0</td><td>72.0</td><td>48.5</td><td>48.5</td></tr>
<tr><td>Port</td><td>[pP]ort[=: |:|=|: |\s/]*(\d1,5)</td><td>96.0</td><td>96.2</td><td>8.1</td><td>8.1</td><td>15.0</td><td>15.0</td></tr>
<tr><td>Configuration</td><td>size\s+(\d+)</td><td>19.2</td><td>19.2</td><td>14.2</td><td>14.2</td><td>16.3</td><td>16.3</td></tr>
</table>

Now

Now

We introduce

Now

We introduce

SDLog

Now

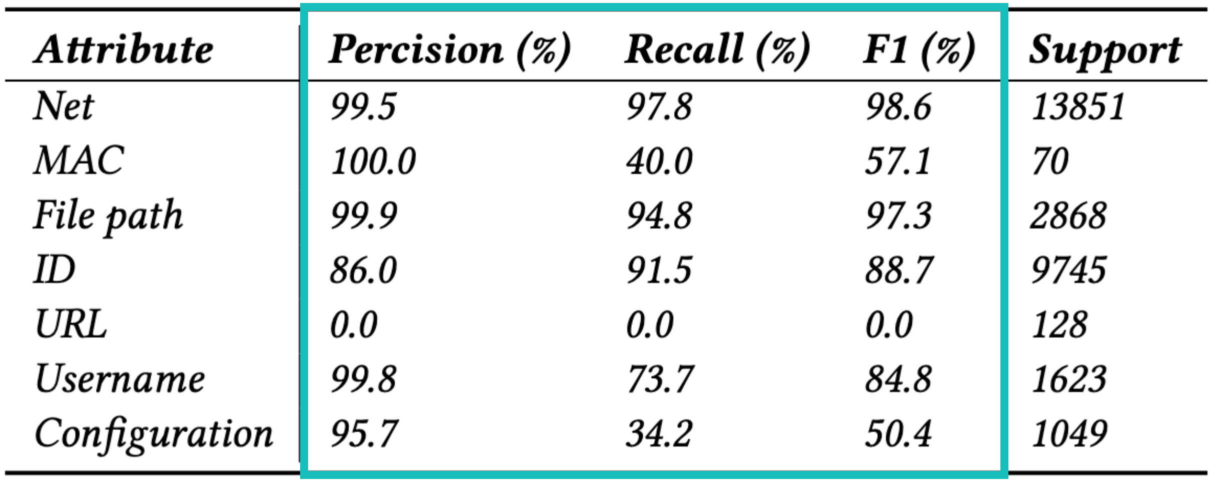
We introduce

SDLog

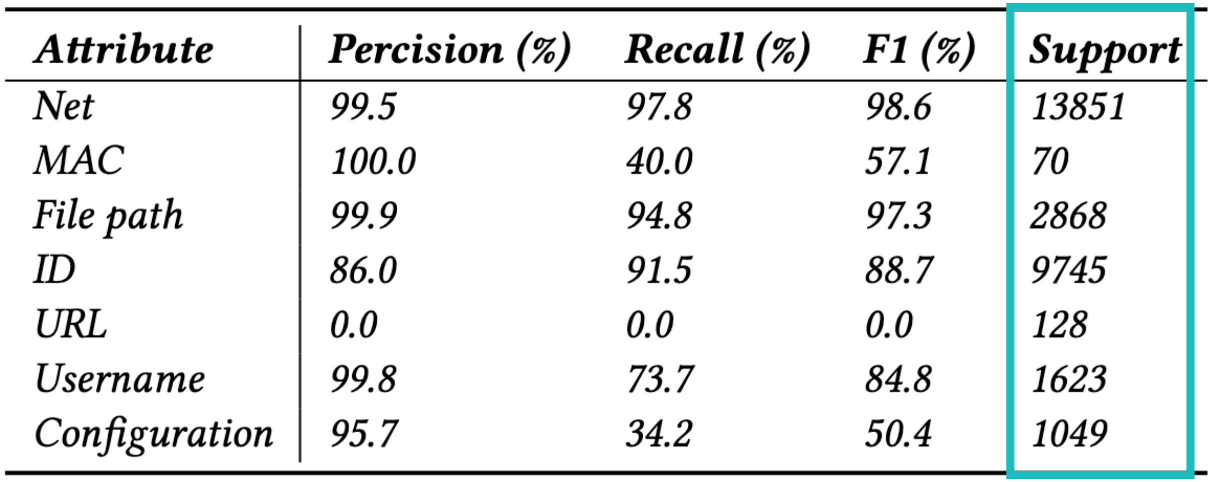
Sensitivity Detector in Logs

Let’s check its results.

Attribute	*Percision (%)*	*Recall (%)*	*F1 (%)*	*Support*
Net	*99.5*	*97.8*	*98.6*	*13851*
MAC	*100.0*	*40.0*	*57.1*	*70*
File path	*99.9*	*94.8*	*97.3*	*2868*
ID	*86.0*	*91.5*	*88.7*	*9745*
URL	*0.0*	*0.0*	*0.0*	*128*
Username	*99.8*	*73.7*	*84.8*	*1623*
Configuration	*95.7*	*34.2*	*50.4*	*1049*

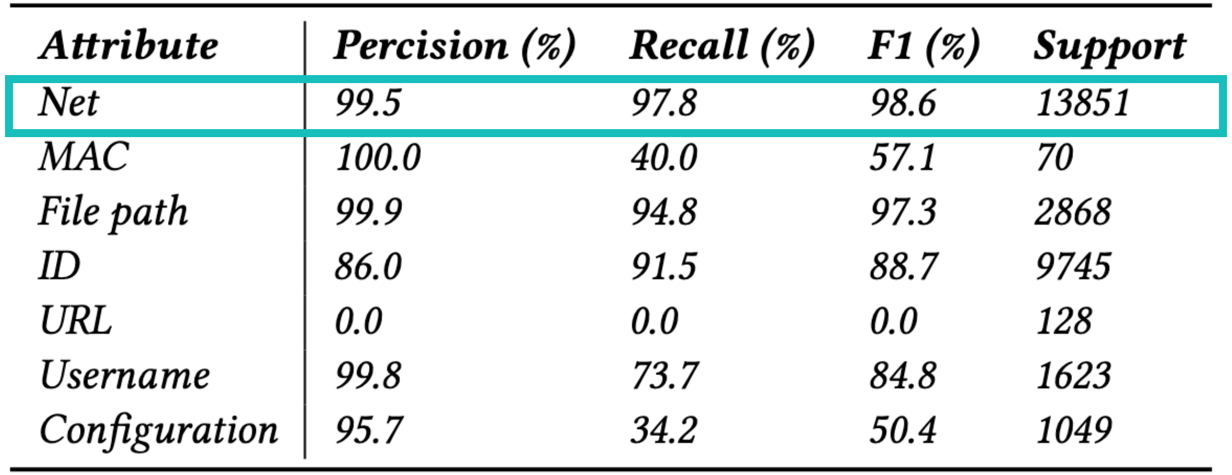


Attribute	*Percision (%)*	*Recall (%)*	*F1 (%)*	*Support*
Net	*99.5*	*97.8*	*98.6*	*13851*
MAC	*100.0*	*40.0*	*57.1*	*70*
File path	*99.9*	*94.8*	*97.3*	*2868*
ID	*86.0*	*91.5*	*88.7*	*9745*
URL	*0.0*	*0.0*	*0.0*	*128*
Username	*99.8*	*73.7*	*84.8*	*1623*
Configuration	*95.7*	*34.2*	*50.4*	*1049*



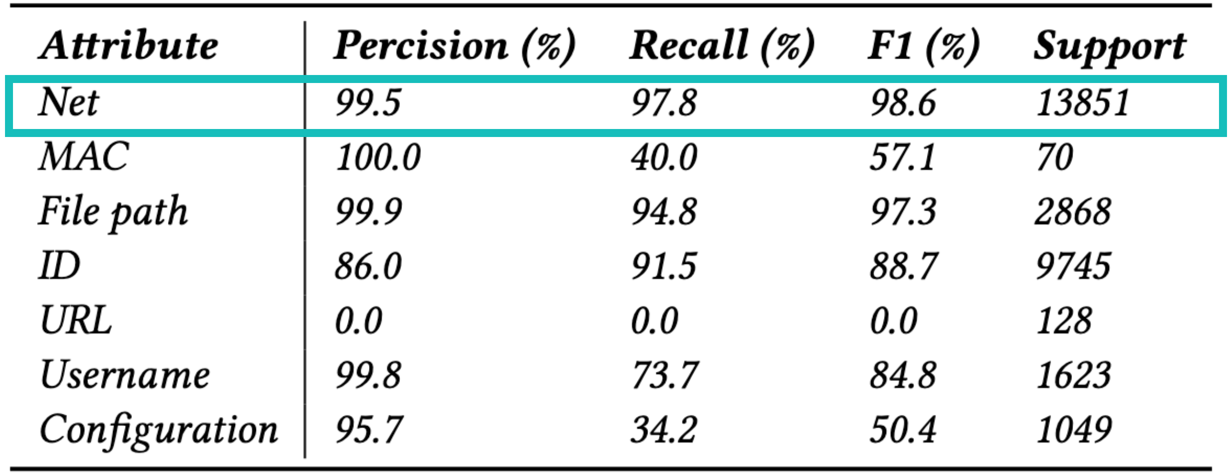
Attribute	*Percision (%)*	*Recall (%)*	*F1 (%)*	*Support*
Net	*99.5*	*97.8*	*98.6*	*13851*
MAC	*100.0*	*40.0*	*57.1*	*70*
File path	*99.9*	*94.8*	*97.3*	*2868*
ID	*86.0*	*91.5*	*88.7*	*9745*
URL	*0.0*	*0.0*	*0.0*	*128*
Username	*99.8*	*73.7*	*84.8*	*1623*
Configuration	*95.7*	*34.2*	*50.4*	*1049*

Net is the combination of
IP address,
port number, and
host name.

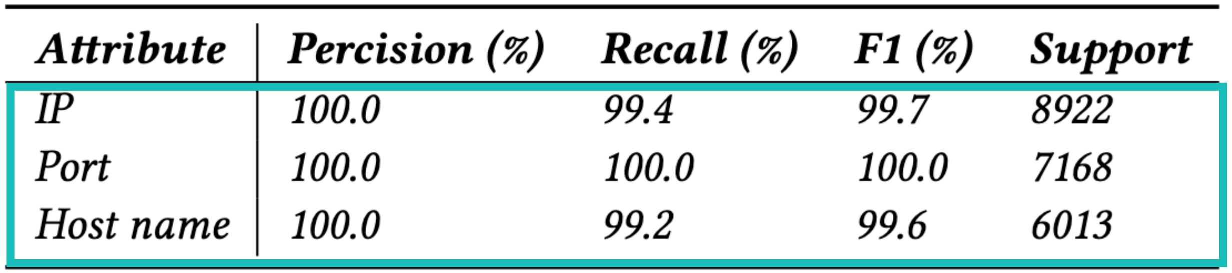


Attribute	*Percision (%)*	*Recall (%)*	*F1 (%)*	*Support*
Net	*99.5*	*97.8*	*98.6*	*13851*
MAC	*100.0*	*40.0*	*57.1*	*70*
File path	*99.9*	*94.8*	*97.3*	*2868*
ID	*86.0*	*91.5*	*88.7*	*9745*
URL	*0.0*	*0.0*	*0.0*	*128*
Username	*99.8*	*73.7*	*84.8*	*1623*
Configuration	*95.7*	*34.2*	*50.4*	*1049*

Net is the combination of
IP address,
port number, and
host name.



Attribute	*Percision (%)*	*Recall (%)*	*F1 (%)*	*Support*
Net	*99.5*	*97.8*	*98.6*	*13851*
MAC	*100.0*	*40.0*	*57.1*	*70*
File path	*99.9*	*94.8*	*97.3*	*2868*
ID	*86.0*	*91.5*	*88.7*	*9745*
URL	*0.0*	*0.0*	*0.0*	*128*
Username	*99.8*	*73.7*	*84.8*	*1623*
Configuration	*95.7*	*34.2*	*50.4*	*1049*



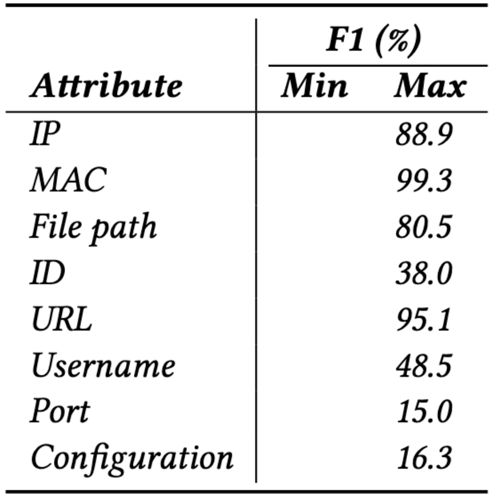
Attribute	*Percision (%)*	*Recall (%)*	*F1 (%)*	*Support*
IP	*100.0*	*99.4*	*99.7*	*8922*
Port	*100.0*	*100.0*	*100.0*	*7168*
Host name	*100.0*	*99.2*	*99.6*	*6013*

First, let’s talk about LIMITATIONS!

Attribute	*Percision (%)*	*Recall (%)*	*F1 (%)*	*Support*
Net	*99.5*	*97.8*	*98.6*	*13851*
MAC	*100.0*	*40.0*	*57.1*	*70*
File path	*99.9*	*94.8*	*97.3*	*2868*
ID	*86.0*	*91.5*	*88.7*	*9745*
URL	*0.0*	*0.0*	*0.0*	*128*
Username	*99.8*	*73.7*	*84.8*	*1623*
Configuration	*95.7*	*34.2*	*50.4*	*1049*

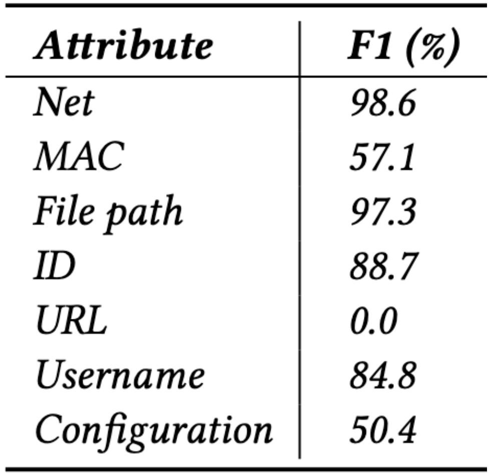
Now, time for COMPARISON!

Best regex patterns

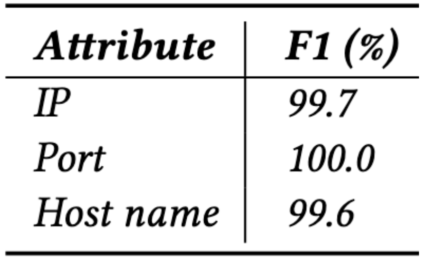


Attribute	F1 (%)	
	Min	Max
IP		88.9
MAC		99.3
File path		80.5
ID		38.0
URL		95.1
Username		48.5
Port		15.0
Configuration		16.3

SDLog

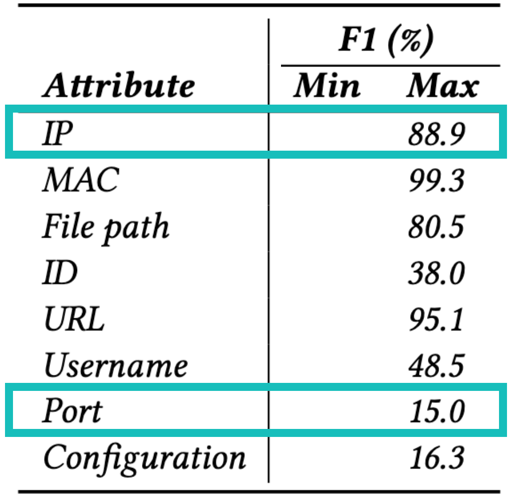


Attribute	F1 (%)
Net	98.6
MAC	57.1
File path	97.3
ID	88.7
URL	0.0
Username	84.8
Configuration	50.4



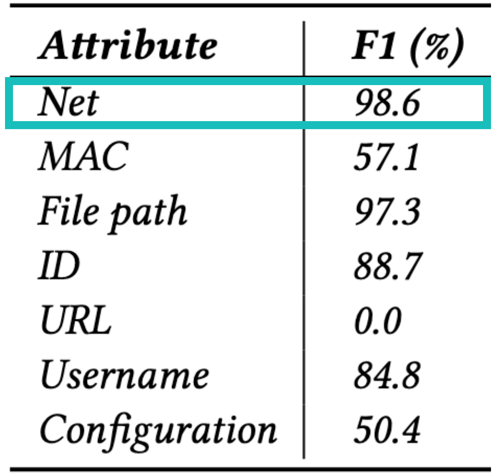
Attribute	F1 (%)
IP	99.7
Port	100.0
Host name	99.6

Best regex patterns

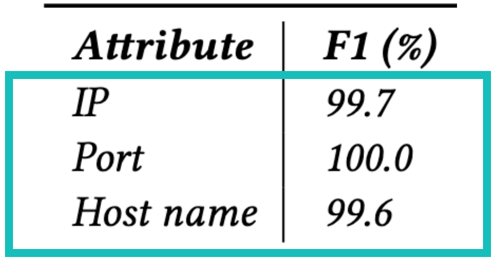


Attribute	*F1 (%)*	
	Min	*Max*
IP		*88.9*
MAC		*99.3*
File path		*80.5*
ID		*38.0*
URL		*95.1*
Username		*48.5*
Port		*15.0*
Configuration		*16.3*

SDLog

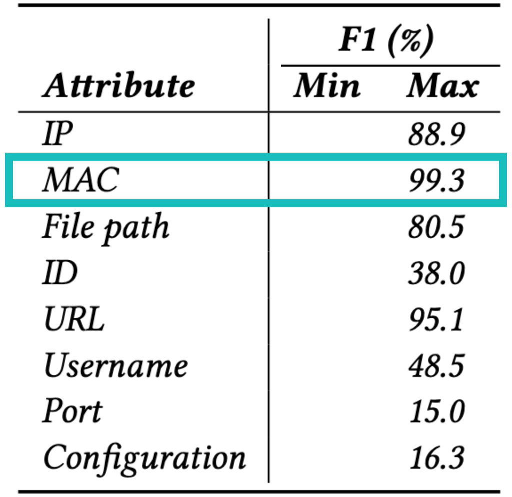


Attribute	*F1 (%)*
Net	*98.6*
MAC	*57.1*
File path	*97.3*
ID	*88.7*
URL	*0.0*
Username	*84.8*
Configuration	*50.4*



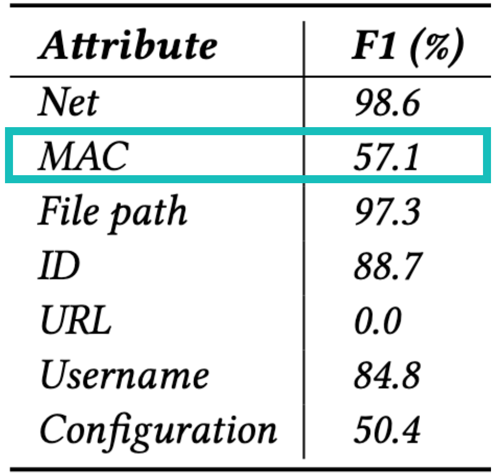
Attribute	*F1 (%)*
IP	*99.7*
Port	*100.0*
Host name	*99.6*

Best regex patterns

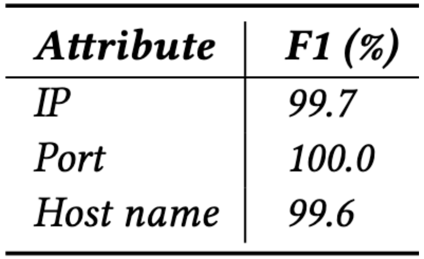


Attribute	F1 (%)	
	Min	Max
IP		88.9
MAC		99.3
File path		80.5
ID		38.0
URL		95.1
Username		48.5
Port		15.0
Configuration		16.3

SDLog

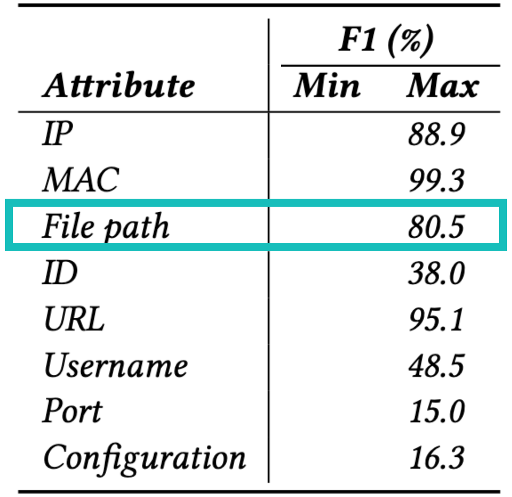


Attribute	F1 (%)
Net	98.6
MAC	57.1
File path	97.3
ID	88.7
URL	0.0
Username	84.8
Configuration	50.4



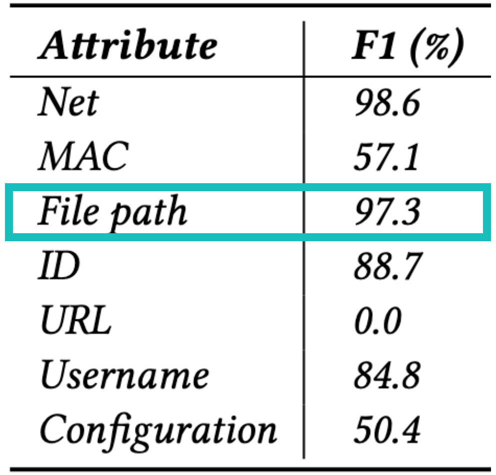
Attribute	F1 (%)
IP	99.7
Port	100.0
Host name	99.6

Best regex patterns

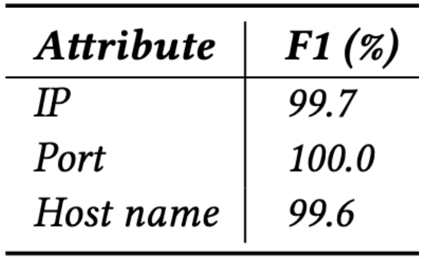


Attribute	F1 (%)	
	Min	Max
IP		88.9
MAC		99.3
File path		80.5
ID		38.0
URL		95.1
Username		48.5
Port		15.0
Configuration		16.3

SDLog

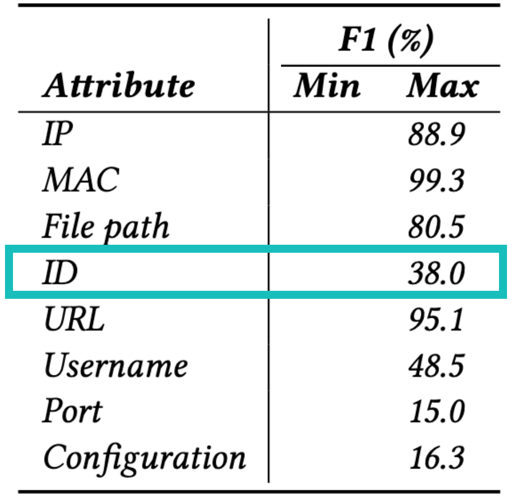


Attribute	F1 (%)
Net	98.6
MAC	57.1
File path	97.3
ID	88.7
URL	0.0
Username	84.8
Configuration	50.4



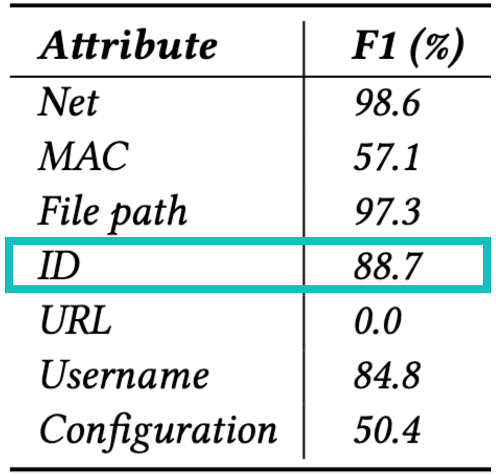
Attribute	F1 (%)
IP	99.7
Port	100.0
Host name	99.6

Best regex patterns

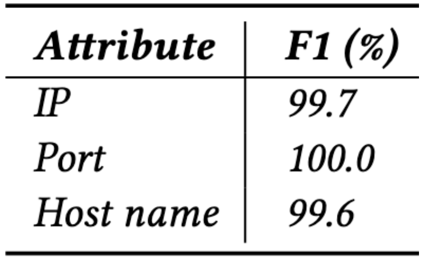


Attribute	*F1 (%)*	
	Min	*Max*
IP		*88.9*
MAC		*99.3*
File path		*80.5*
ID		*38.0*
URL		*95.1*
Username		*48.5*
Port		*15.0*
Configuration		*16.3*

SDLog

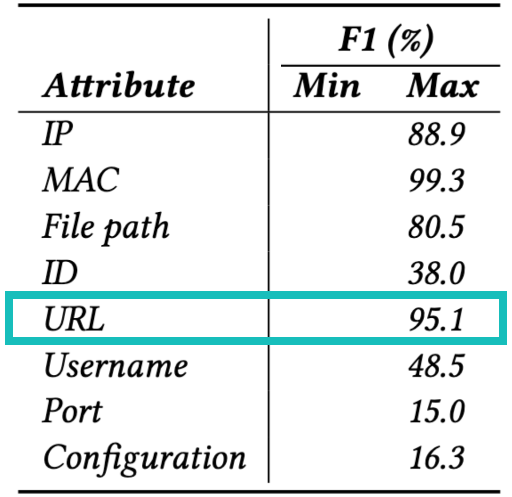


Attribute	*F1 (%)*
Net	*98.6*
MAC	*57.1*
File path	*97.3*
ID	*88.7*
URL	*0.0*
Username	*84.8*
Configuration	*50.4*



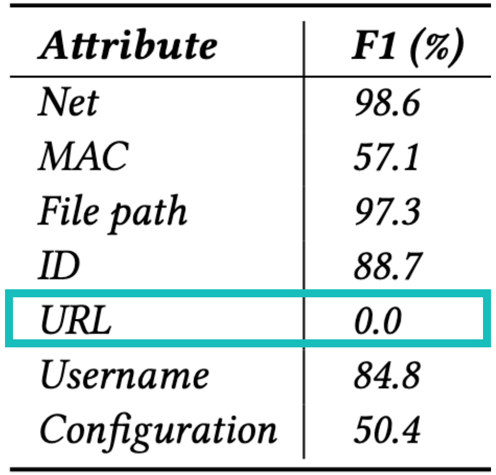
Attribute	*F1 (%)*
IP	*99.7*
Port	*100.0*
Host name	*99.6*

Best regex patterns

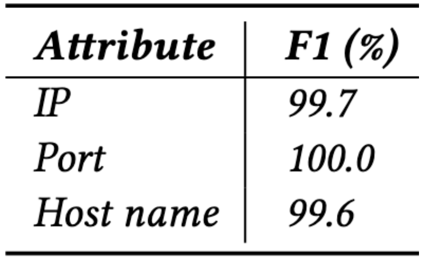


Attribute	F1 (%)	
	Min	Max
IP		88.9
MAC		99.3
File path		80.5
ID		38.0
URL		95.1
Username		48.5
Port		15.0
Configuration		16.3

SDLog

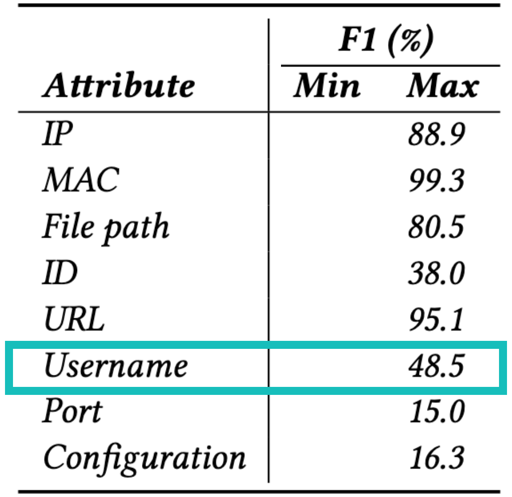


Attribute	F1 (%)
Net	98.6
MAC	57.1
File path	97.3
ID	88.7
URL	0.0
Username	84.8
Configuration	50.4



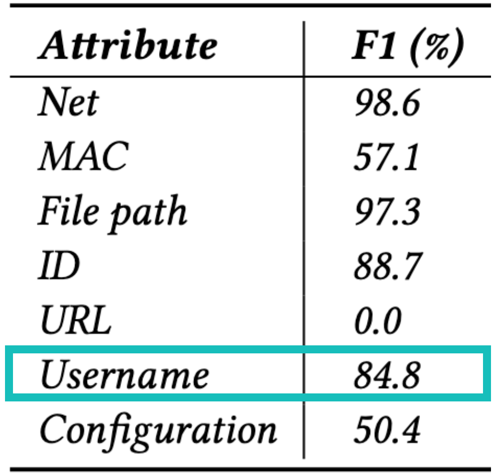
Attribute	F1 (%)
IP	99.7
Port	100.0
Host name	99.6

Best regex patterns

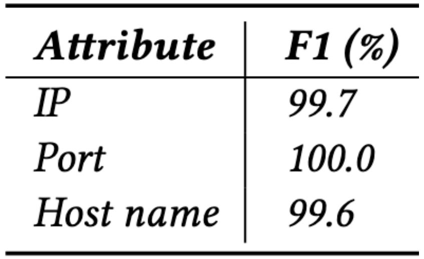


Attribute	F1 (%)	
	Min	Max
IP		88.9
MAC		99.3
File path		80.5
ID		38.0
URL		95.1
Username		48.5
Port		15.0
Configuration		16.3

SDLog

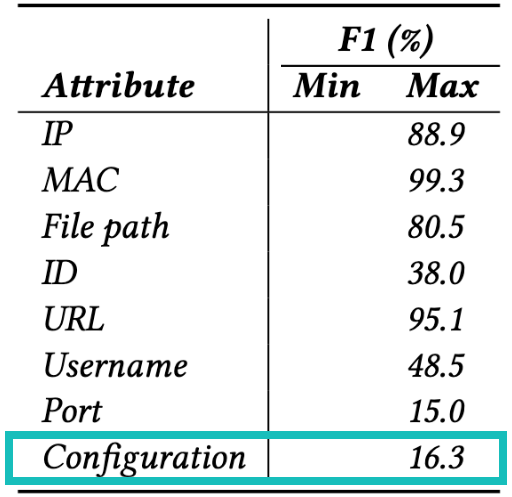


Attribute	F1 (%)
Net	98.6
MAC	57.1
File path	97.3
ID	88.7
URL	0.0
Username	84.8
Configuration	50.4



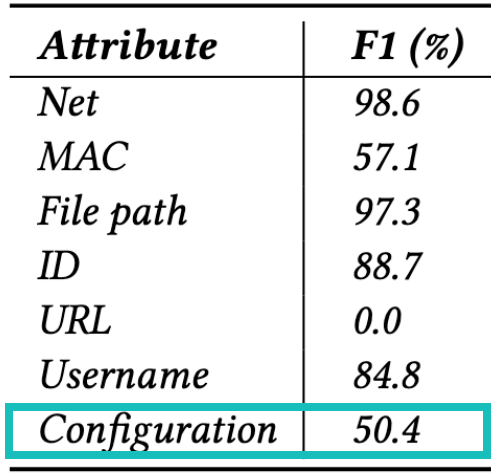
Attribute	F1 (%)
IP	99.7
Port	100.0
Host name	99.6

Best regex patterns

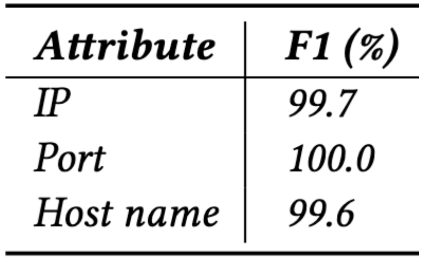


Attribute	F1 (%)	
	Min	Max
IP		88.9
MAC		99.3
File path		80.5
ID		38.0
URL		95.1
Username		48.5
Port		15.0
Configuration		16.3

SDLog



Attribute	F1 (%)
Net	98.6
MAC	57.1
File path	97.3
ID	88.7
URL	0.0
Username	84.8
Configuration	50.4



Attribute	F1 (%)
IP	99.7
Port	100.0
Host name	99.6

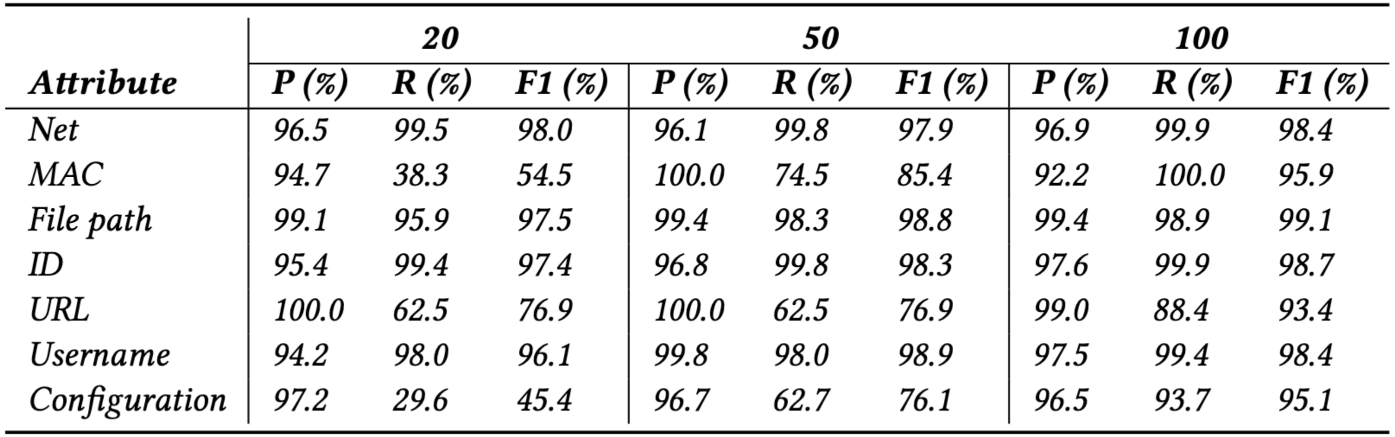
Now, you might ask:

“Is there any way to improve the performance of SDLog?”

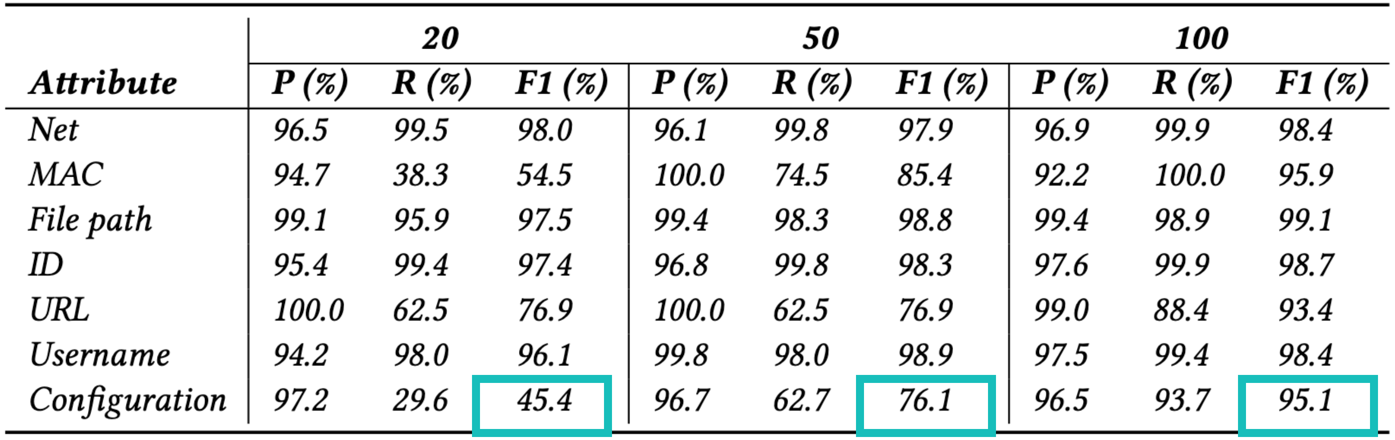
YES!

Organizations can fine-tune SDLog with their datasets.

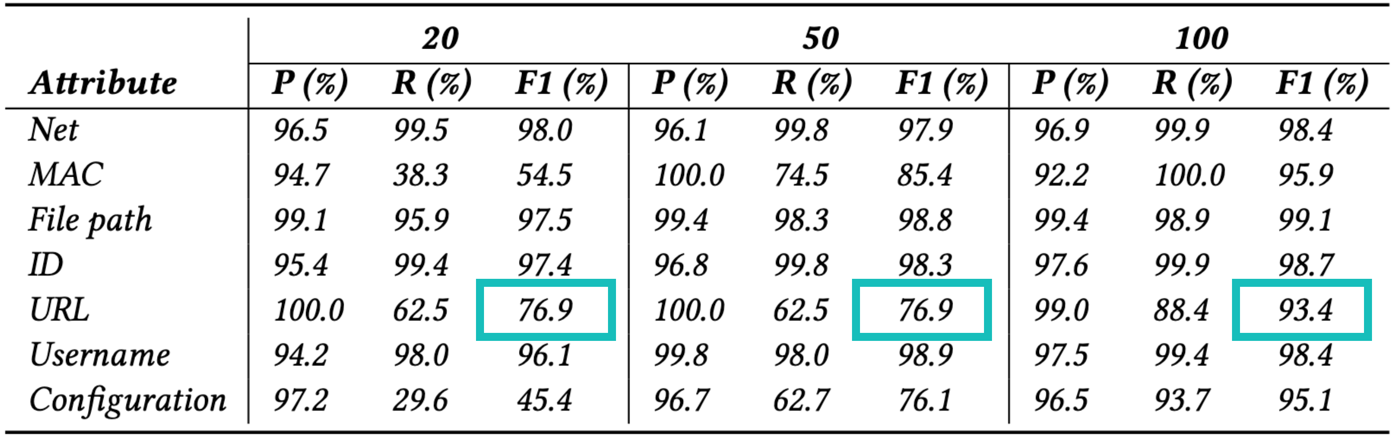
Let’s see the results!



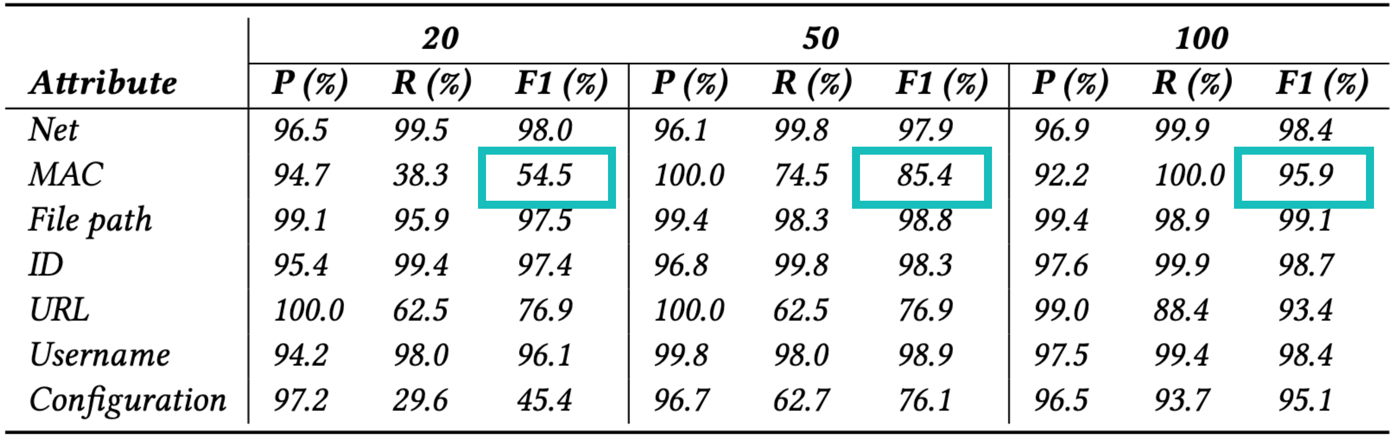
Attribute	***20***			***50***			***100***		
	P (%)	***R (%)***	***F1 (%)***	***P (%)***	***R (%)***	***F1 (%)***	***P (%)***	***R (%)***	***F1 (%)***
Net	*96.5*	*99.5*	*98.0*	*96.1*	*99.8*	*97.9*	*96.9*	*99.9*	*98.4*
MAC	*94.7*	*38.3*	*54.5*	*100.0*	*74.5*	*85.4*	*92.2*	*100.0*	*95.9*
File path	*99.1*	*95.9*	*97.5*	*99.4*	*98.3*	*98.8*	*99.4*	*98.9*	*99.1*
ID	*95.4*	*99.4*	*97.4*	*96.8*	*99.8*	*98.3*	*97.6*	*99.9*	*98.7*
URL	*100.0*	*62.5*	*76.9*	*100.0*	*62.5*	*76.9*	*99.0*	*88.4*	*93.4*
Username	*94.2*	*98.0*	*96.1*	*99.8*	*98.0*	*98.9*	*97.5*	*99.4*	*98.4*
Configuration	*97.2*	*29.6*	*45.4*	*96.7*	*62.7*	*76.1*	*96.5*	*93.7*	*95.1*



Attribute	*20*			*50*			*100*		
	P (%)	*R (%)*	*F1 (%)*	*P (%)*	*R (%)*	*F1 (%)*	*P (%)*	*R (%)*	*F1 (%)*
Net	*96.5*	*99.5*	*98.0*	*96.1*	*99.8*	*97.9*	*96.9*	*99.9*	*98.4*
MAC	*94.7*	*38.3*	*54.5*	*100.0*	*74.5*	*85.4*	*92.2*	*100.0*	*95.9*
File path	*99.1*	*95.9*	*97.5*	*99.4*	*98.3*	*98.8*	*99.4*	*98.9*	*99.1*
ID	*95.4*	*99.4*	*97.4*	*96.8*	*99.8*	*98.3*	*97.6*	*99.9*	*98.7*
URL	*100.0*	*62.5*	*76.9*	*100.0*	*62.5*	*76.9*	*99.0*	*88.4*	*93.4*
Username	*94.2*	*98.0*	*96.1*	*99.8*	*98.0*	*98.9*	*97.5*	*99.4*	*98.4*
Configuration	*97.2*	*29.6*	*45.4*	*96.7*	*62.7*	*76.1*	*96.5*	*93.7*	*95.1*



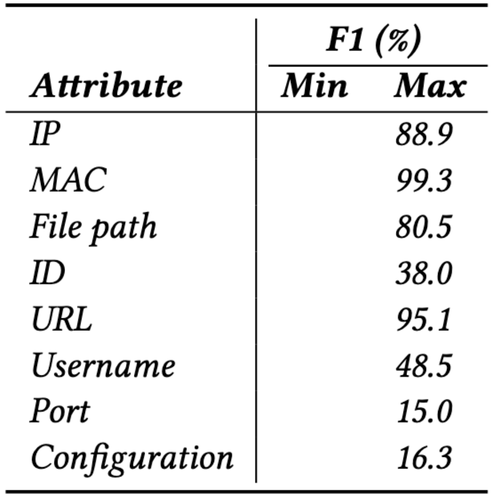
Attribute	20			50			100		
	P (%)	**R (%)**	**F1 (%)**	**P (%)**	**R (%)**	**F1 (%)**	**P (%)**	**R (%)**	**F1 (%)**
Net	*96.5*	*99.5*	*98.0*	*96.1*	*99.8*	*97.9*	*96.9*	*99.9*	*98.4*
MAC	*94.7*	*38.3*	*54.5*	*100.0*	*74.5*	*85.4*	*92.2*	*100.0*	*95.9*
File path	*99.1*	*95.9*	*97.5*	*99.4*	*98.3*	*98.8*	*99.4*	*98.9*	*99.1*
ID	*95.4*	*99.4*	*97.4*	*96.8*	*99.8*	*98.3*	*97.6*	*99.9*	*98.7*
URL	*100.0*	*62.5*	*76.9*	*100.0*	*62.5*	*76.9*	*99.0*	*88.4*	*93.4*
Username	*94.2*	*98.0*	*96.1*	*99.8*	*98.0*	*98.9*	*97.5*	*99.4*	*98.4*
Configuration	*97.2*	*29.6*	*45.4*	*96.7*	*62.7*	*76.1*	*96.5*	*93.7*	*95.1*



Attribute	**20**			**50**			**100**		
	P (%)	**R (%)**	**F1 (%)**	**P (%)**	**R (%)**	**F1 (%)**	**P (%)**	**R (%)**	**F1 (%)**
Net	*96.5*	*99.5*	*98.0*	*96.1*	*99.8*	*97.9*	*96.9*	*99.9*	*98.4*
MAC	*94.7*	*38.3*	*54.5*	*100.0*	*74.5*	*85.4*	*92.2*	*100.0*	*95.9*
File path	*99.1*	*95.9*	*97.5*	*99.4*	*98.3*	*98.8*	*99.4*	*98.9*	*99.1*
ID	*95.4*	*99.4*	*97.4*	*96.8*	*99.8*	*98.3*	*97.6*	*99.9*	*98.7*
URL	*100.0*	*62.5*	*76.9*	*100.0*	*62.5*	*76.9*	*99.0*	*88.4*	*93.4*
Username	*94.2*	*98.0*	*96.1*	*99.8*	*98.0*	*98.9*	*97.5*	*99.4*	*98.4*
Configuration	*97.2*	*29.6*	*45.4*	*96.7*	*62.7*	*76.1*	*96.5*	*93.7*	*95.1*

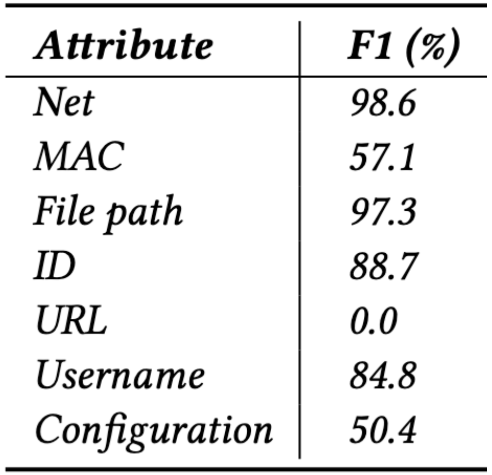
Let’s compare them again!

Best regex patterns



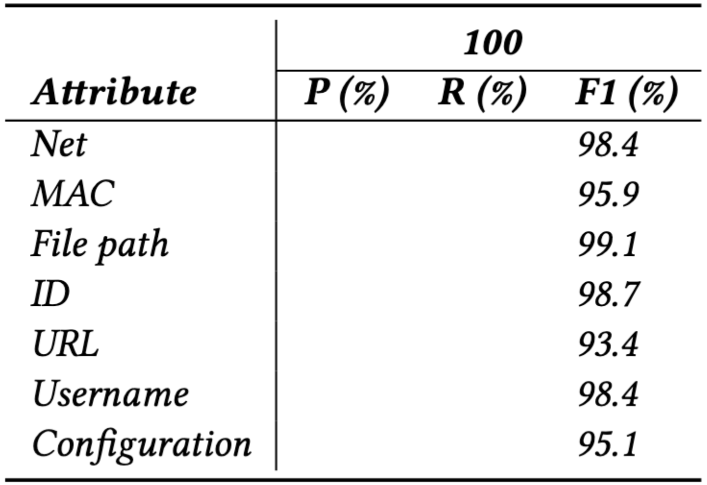
Attribute	*F1 (%)*	
	Min	*Max*
IP		*88.9*
MAC		*99.3*
File path		*80.5*
ID		*38.0*
URL		*95.1*
Username		*48.5*
Port		*15.0*
Configuration		*16.3*

SDLog



Attribute	*F1 (%)*
Net	*98.6*
MAC	*57.1*
File path	*97.3*
ID	*88.7*
URL	*0.0*
Username	*84.8*
Configuration	*50.4*

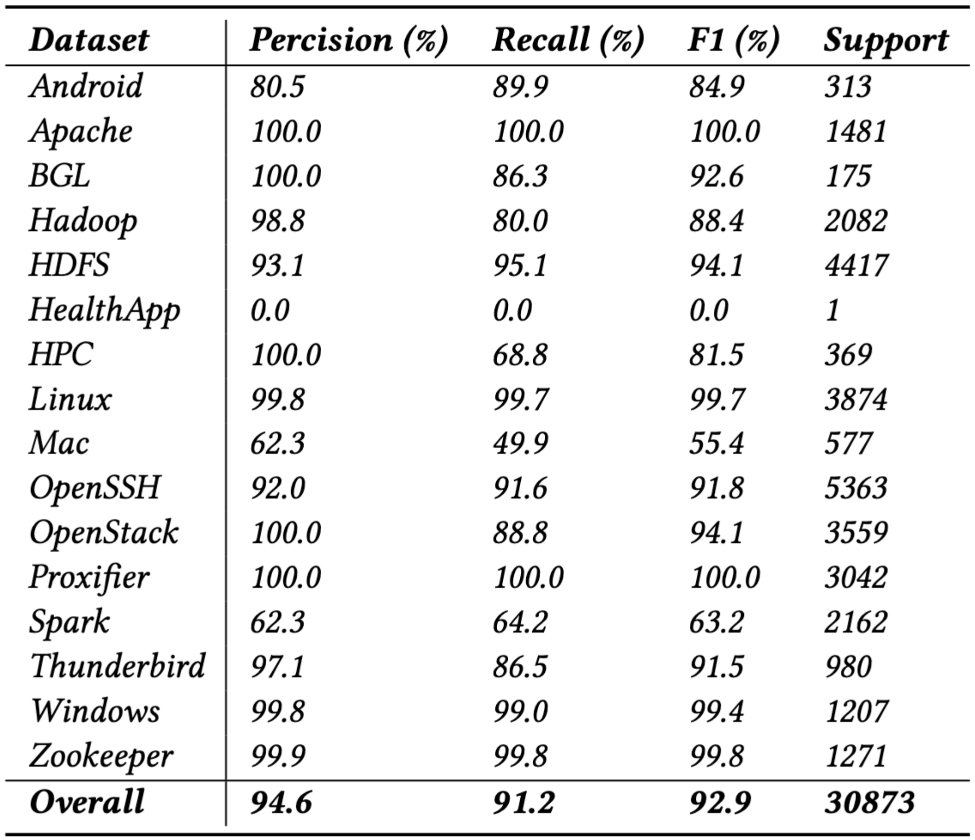
Fine-tuned SDLog



Attribute	*100*		
	P (%)	*R (%)*	*F1 (%)*
Net			*98.4*
MAC			*95.9*
File path			*99.1*
ID			*98.7*
URL			*93.4*
Username			*98.4*
Configuration			*95.1*

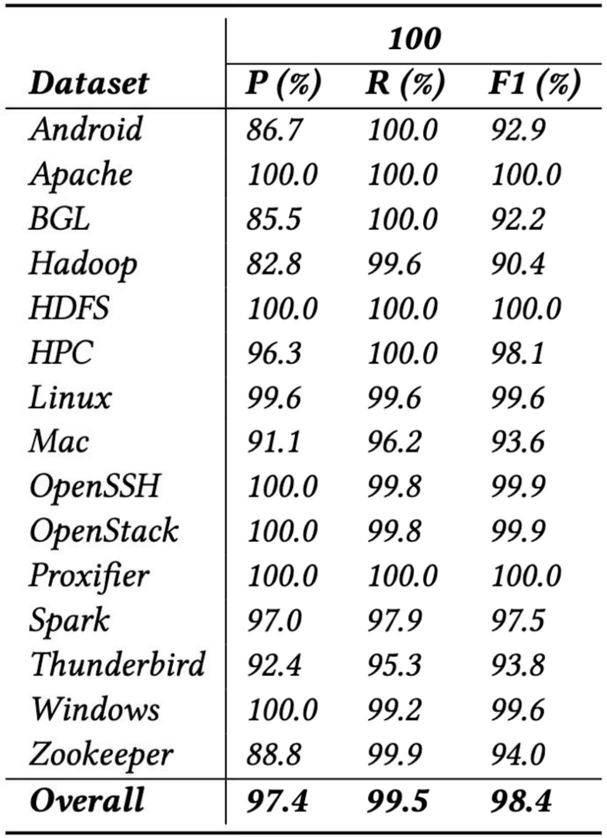
So, how much of the sensitive information can SDLog detect?

SDLog



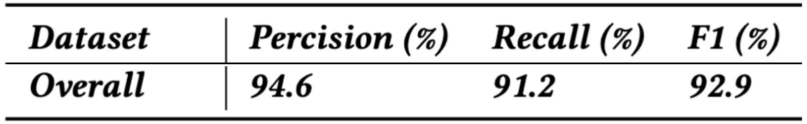
Dataset	Percision (%)	Recall (%)	F1 (%)	Support
Android	80.5	89.9	84.9	313
Apache	100.0	100.0	100.0	1481
BGL	100.0	86.3	92.6	175
Hadoop	98.8	80.0	88.4	2082
HDFS	93.1	95.1	94.1	4417
HealthApp	0.0	0.0	0.0	1
HPC	100.0	68.8	81.5	369
Linux	99.8	99.7	99.7	3874
Mac	62.3	49.9	55.4	577
OpenSSH	92.0	91.6	91.8	5363
OpenStack	100.0	88.8	94.1	3559
Proxifier	100.0	100.0	100.0	3042
Spark	62.3	64.2	63.2	2162
Thunderbird	97.1	86.5	91.5	980
Windows	99.8	99.0	99.4	1207
Zookeeper	99.9	99.8	99.8	1271
Overall	**94.6**	**91.2**	**92.9**	**30873**

Fine-tuned SDLog



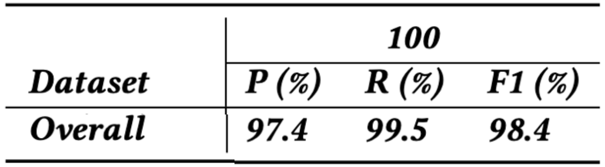
	100		
Dataset	**P (%)**	**R (%)**	**F1 (%)**
Android	86.7	100.0	92.9
Apache	100.0	100.0	100.0
BGL	85.5	100.0	92.2
Hadoop	82.8	99.6	90.4
HDFS	100.0	100.0	100.0
HPC	96.3	100.0	98.1
Linux	99.6	99.6	99.6
Mac	91.1	96.2	93.6
OpenSSH	100.0	99.8	99.9
OpenStack	100.0	99.8	99.9
Proxifier	100.0	100.0	100.0
Spark	97.0	97.9	97.5
Thunderbird	92.4	95.3	93.8
Windows	100.0	99.2	99.6
Zookeeper	88.8	99.9	94.0
Overall	**97.4**	**99.5**	**98.4**

SDLog



Dataset	*Percision (%)*	*Recall (%)*	*F1 (%)*
Overall	*94.6*	*91.2*	*92.9*

Fine-tuned SDLog



	100		
Dataset	*P (%)*	*R (%)*	*F1 (%)*
Overall	*97.4*	*99.5*	*98.4*

How did we build it?

This is BERT.

It was introduced by Google in 2018.

BERT was trained on ~3.3B words and has ~110M parameters.

This is CodeBERT.

It was introduced by Microsoft in 2020.

CodeBERT was trained on ~8.5M codes and has ~125M parameters.

This is SDLog.

SDLog was introduced by Aghili et al. in 2025!

SDLog uses CodeBERT as backbone and is fine-tuned with 32,000 software logs.

Now, if you are still interested

And have questions like:

“How much time does it take to run SDLog?”

“How much time does it take to run SDLog?”

“How complex is it to fine-tune SDLog?”

“How much time does it take to run SDLog?”

“How complex is it to fine-tune SDLog?”

“How much GPU do I need to fine-tune it?”

We have good news for you!

It only takes several minutes to run SDLog.

It only takes several minutes to fine-tine SDLog with 100 samples of your dataset.

It only takes 2-3 days to label 100 samples.

And the best part is…

We have shared all our models and scripts.

You can find our replication package here:

github.com/mooselab/SDLog

Someone from IT
sends you a message:

Someone from IT
sends you a message:

IP address
MAC address
Host name
File path
ID
URL
Username
Port number
Configuration details

Someone from IT sends you a message:

IP address
MAC address
Host name
File path
ID
URL
Username
Port number
Configuration details

Then we use this regular expression:

```
\d+\.\d+\.\d+\.\d+
```

Someone from IT sends you a message:

IP address
MAC address
Host name
File path
ID
URL
Username
Port number
Configuration details

Then we use this regular expression:

```
\d+\.\d+\.\d+\.\d+
```

There is NO common ground truth for regular expressions!

Someone from IT
sends you a message:

There is NO common
ground truth for
regular expressions!

IP address
MAC address
Host name
File path
ID
URL
Username
Port number
Configuration details

Now
We introduce

SDLog

Sensitivity Detector in Logs

Then we use this
regular expression:
`\d+\.\d+\.\d+\.\d+`

Someone from IT
sends you a message:

IP address
MAC address
Host name
File path
ID
URL
Username
Port number
Configuration details

Then we use this
regular expression:
`\d+\.\d+\.\d+\.\d+`

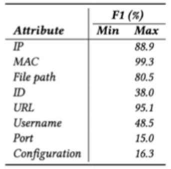
There is NO common
ground truth for
regular expressions!

Now
We introduce

SDLog

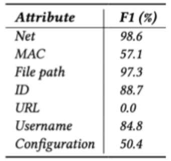
Sensitivity Detector in Logs

Best regex patterns



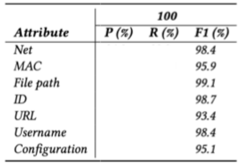
Attribute	F1 (%)	
	Min	Max
IP		88.9
MAC		99.3
File path		80.5
ID		38.0
URL		95.1
Username		48.5
Port		15.0
Configuration		16.3

SDLog



Attribute	F1 (%)
Net	98.6
MAC	57.1
File path	97.3
ID	88.7
URL	0.0
Username	84.8
Configuration	50.4

Fine-tuned SDLog



Attribute	100		
	P (%)	R (%)	F1 (%)
Net			98.4
MAC			95.9
File path			99.1
ID			98.7
URL			93.4
Username			98.4
Configuration			95.1

Someone from IT
sends you a message:

There is NO common
ground truth for
regular expressions!

IP address
MAC address
Host name
File path
ID
URL
Username
Port number
Configuration details

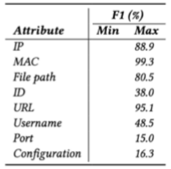
Now
We introduce

SDLog

Sensitivity Detector in Logs

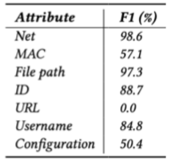
Then we use this
regular expression:
\d+\.\d+\.\d+\.\d+

Best regex patterns



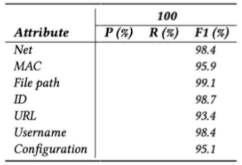
Attribute	F1 (%)	
	Min	Max
IP		88.9
MAC		99.3
File path		80.5
ID		38.0
URL		95.1
Username		48.5
Port		15.0
Configuration		16.3

SDLog



Attribute	F1 (%)
Net	98.6
MAC	57.1
File path	97.3
ID	88.7
URL	0.0
Username	84.8
Configuration	50.4

Fine-tuned SDLog



Attribute	100		
	P (%)	R (%)	F1 (%)
Net			98.4
MAC			95.9
File path			99.1
ID			98.7
URL			93.4
Username			98.4
Configuration			95.1

SDLog uses
CodeBERT as
backbone and is
fine-tuned with
32,000 software
logs.

Someone from IT
sends you a message:

There is NO common
ground truth for
regular expressions!

SDLog uses CodeBERT as backbone and is fine-tuned with 32,000 software logs.

IP address
MAC address
Host name
File path
ID
URL
Username
Port number
Configuration details

Now
We introduce

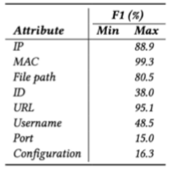
SDLog

Sensitivity Detector in Logs

It only takes several minutes to run SDLog.

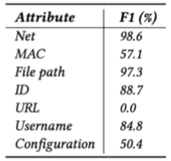
Then we use this
regular expression:
`\d+\.\d+\.\d+\.\d+`

Best regex patterns



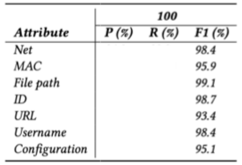
	F1 (%)	
Attribute	Min	Max
IP		88.9
MAC		99.3
File path		80.5
ID		38.0
URL		95.1
Username		48.5
Port		15.0
Configuration		16.3

SDLog



Attribute	F1 (%)
Net	98.6
MAC	57.1
File path	97.3
ID	88.7
URL	0.0
Username	84.8
Configuration	50.4

Fine-tuned SDLog



	100		
Attribute	P (%)	R (%)	F1 (%)
Net			98.4
MAC			95.9
File path			99.1
ID			98.7
URL			93.4
Username			98.4
Configuration			95.1

Someone from IT
sends you a message:

There is NO common
ground truth for
regular expressions!

SDLog uses
CodeBERT as
backbone and is
fine-tuned with
32,000 software
logs.

IP address
MAC address
Host name
File path
ID
URL
Username
Port number
Configuration details

Now
We introduce

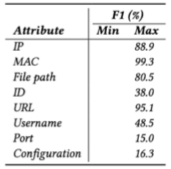
SDLog

Sensitivity Detector in Logs

It only takes several minutes
to run SDLog.

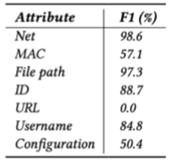
Then we use this
regular expression:
`\d+\.\d+\.\d+\.\d+`

Best regex patterns



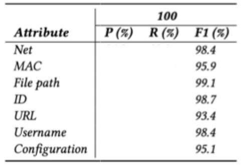
Attribute	*F1 (%)*	
	Min	*Max*
IP		*88.9*
MAC		*99.3*
File path		*80.5*
ID		*38.0*
URL		*95.1*
Username		*48.5*
Port		*15.0*
Configuration		*16.3*

SDLog



Attribute	*F1 (%)*
Net	*98.6*
MAC	*57.1*
File path	*97.3*
ID	*88.7*
URL	*0.0*
Username	*84.8*
Configuration	*50.4*

Fine-tuned SDLog



Attribute	*100*		
	P (%)	*R (%)*	*F1 (%)*
Net			*98.4*
MAC			*95.9*
File path			*99.1*
ID			*98.7*
URL			*93.4*
Username			*98.4*
Configuration			*95.1*

You can find our
replication package here:

github.com/mooselab/SDLog