Supplementary material

S1 Synthetic fixtures: which instrument treats which construct as hidden

Each fixture is a small HTML message with one visible sentence and one target sentence. “hidden” means that the instrument flags the target as hidden or omits it from the text it keeps as visible. The last four rows are plainly visible markup. No mailbox data are involved.

Fixture	This study (static analyser)	This study (Chromium)	Script of Guo et al.	SpamAssassin text	Rspamd text	Rspamd symbols	SpamAssassin rules and sub-rules
display_none_inline	hidden	hidden	hidden	hidden	hidden	MANY_INVISIBLE_PARTS	__STY_INVIS, __STY_INVIS_1, __STY_INVIS_1_MINFP
display_none_class	hidden	hidden	visible	visible	hidden	MANY_INVISIBLE_PARTS	
display_none_ancestor	hidden	hidden	visible	hidden	hidden	MANY_INVISIBLE_PARTS	__STY_INVIS, __STY_INVIS_1, __STY_INVIS_1_MINFP
hidden_attribute	hidden	hidden	visible	visible	hidden	MANY_INVISIBLE_PARTS	
visibility_hidden	hidden	hidden	hidden	hidden	hidden	MANY_INVISIBLE_PARTS	__STY_INVIS, __STY_INVIS_1, __STY_INVIS_1_MINFP
opacity_0	hidden	hidden	hidden	visible	hidden	MANY_INVISIBLE_PARTS	
font_size_0	hidden	hidden	visible	visible	hidden	MANY_INVISIBLE_PARTS, ZERO_FONT	
font_size_1px	hidden	hidden	hidden	visible	hidden	MANY_INVISIBLE_PARTS	__AC_TINY_FONT
color_transparent	hidden	hidden	visible	hidden	hidden	MANY_INVISIBLE_PARTS, R_WHITE_ON_WHITE	HTML_FONT_LOW_CONTRAST
white_on_white_css	hidden	hidden	hidden	hidden	hidden	MANY_INVISIBLE_PARTS, R_WHITE_ON_WHITE	HTML_FONT_LOW_CONTRAST
white_on_white_font	hidden	hidden	hidden	hidden	hidden	MANY_INVISIBLE_PARTS, R_WHITE_ON_WHITE	HTML_FONT_LOW_CONTRAST
offscreen_left	hidden	hidden	hidden	visible	hidden	MANY_INVISIBLE_PARTS	
text_indent	hidden	hidden	hidden	visible	hidden	MANY_INVISIBLE_PARTS	
max_height_0	hidden	hidden	visible	visible	hidden	MANY_INVISIBLE_PARTS	
clip_rect	hidden	hidden	hidden	visible	hidden	MANY_INVISIBLE_PARTS	
mso_hide	visible	visible	visible	visible	visible		
preheader_idiom	hidden	hidden	hidden	hidden	hidden	MANY_INVISIBLE_PARTS	__STY_INVIS, __STY_INVIS_1, __STY_INVIS_1_MINFP
comment	n/a	n/a	n/a	hidden	hidden		
visible_font_size_1	visible	visible	visible	hidden	visible		
visible_white_on_dark_cell	visible	visible	visible	hidden	visible		HTML_FONT_LOW_CONTRAST
visible_white_on_dark_class	visible	visible	hidden	hidden	visible		HTML_FONT_LOW_CONTRAST
visible_grey_9px	visible	visible	visible	visible	visible		

S2 Rules by year, mailbox 1 (random sample, % of HTML messages)

year	HTML messages	Hidden region 20+	Leading only	Residual	Guo et al. rule	Guo et al. rule, released script	Nine declarations (Talos)	SA __STY_INVIS	SA LOW_CONTRAST	SA invisible buffer 20+	SA MPART_ALT_DIFF	Rspamd MANY_INVISIBLE_PARTS	Rspamd ZERO_FONT	Rspamd HIDDEN_TEXT	Jaccard > 0.2
2005	14	0	0	0	0	0	0	0	85.7	85.7	0	0	0	0	0
2006	125	0	0	0	0	0	2.4	0	73.6	83.2	0	3.2	0	0	0
2007	166	0	0	0	0	0.6	3.6	3	53	71.1	3	2.4	0	3.6	0
2008	163	0	0	0	0	0	0	0	40.5	60.1	0.6	0.6	0	0.6	0
2009	473	0	0	0	0	8.2	0	0	58.8	70.8	0.8	18	0	0.2	0
2010	684	0.1	0.1	0	0	4.1	0.1	0	57.3	65.5	1.9	18.1	0	0.7	0
2011	676	0	0	0	0	1.9	1.5	0.1	56.4	61.7	3.7	10.4	0.4	0.3	0
2012	1017	0.2	0.2	0	0	0.8	16	1.5	47.5	55.1	3	3.8	0.4	0	0
2013	1099	0	0	0	0	1	25	1	48.1	54.1	4.6	6.5	0.1	0.2	0
2014	997	2.9	2.7	0	2.4	2.9	29.3	4.4	39.3	45.4	8.3	9.1	1.8	1.6	0
2015	890	3.5	2.4	0.3	2.8	2.1	28.7	5.7	25.6	32.5	8	10	1.9	0.6	0
2016	877	7.8	5.2	2.3	3	6.8	39.2	14.1	33	32	9.2	13.7	1	1.6	0

year	HTML messages	Hidden region 20+	Leading only	Residual	Guo et al. rule	Guo et al. rule, released script	Nine declarations (Talos)	SA __STY_INVIS	SA LOW_CONTRAST	SA invisible buffer 20+	SA MPART_ALT_DIFF	Rspamd MANY_INVISIBLE_PARTS	Rspamd ZERO_FONT	Rspamd HIDDEN_TEXT	Jaccard > 0.2
2017	1060	14.5	12	2.5	7.1	11.6	45.5	21.1	34.6	33.5	8.6	22.2	2.5	0.8	0.6
2018	1141	10.5	6.7	3.8	4.9	3.2	43.5	17.2	25.9	30.1	10.4	15.3	6.1	1.8	2.8
2019	1235	15.3	11.8	3.2	8.2	3.9	54.1	24.3	28.1	34.5	10.9	19.1	6.6	1.1	4.2
2020	1404	21.2	15.5	4.5	8.8	3.3	65.1	30.2	30.6	36.2	15.4	22.8	8.4	2.6	4.8
2021	1346	20.4	16.4	3.8	8.8	4.2	64	36.5	30.8	35.1	14.1	24.2	7.1	1.6	4.6
2022	1479	22.9	18.8	3.5	8.1	4.2	69.9	39.6	32	36.8	15.7	29.3	8.5	1.8	3.6
2023	1372	25.2	21.6	2.8	8.7	5.5	72.2	42.4	34	37.1	15.4	31.3	7.5	2.3	3.7
2024	1435	25.6	22.2	2.7	9.4	7.7	77.2	44.4	30.8	36.1	17.6	32.4	8.3	3.4	1.8
2025	1519	37.5	33.4	3.4	17.7	15.1	84.6	56	32.5	42.8	16.9	41.5	13.4	3	1.7
2026	1581	33	29.2	3.5	17.8	14	84.1	49.6	31.1	40.1	13.2	44.3	9.7	1.2	1.6

S3 Rules by year, Spam Archive sample (% of HTML messages)

The two filters were run on the years 2021 to 2026 only.

year	HTML messages	Hidden region 20+	Leading only	Residual	Guo et al. rule	Guo et al. rule, released script	Nine declarations (Talos)	Jaccard > 0.2
2000	698	0	0	0	0	0	0.3	0
2001	1444	1.2	0.2	0.8	0.6	0.8	0	0.1
2002	1860	3.3	0.1	3.3	1.1	1.3	0.3	0.2
2003	2324	11.1	1.6	9.5	10.8	7.5	0.3	6
2004	2172	16.9	3.7	13.3	14.9	10.1	4.1	12.7
2005	1600	5	0.2	4.8	3.8	2.1	1.7	3.9
2006	1413	2.2	0	2.2	2.1	2	0.5	1.4
2007	1844	2.8	0.1	2.7	3	2.1	0.2	2.8
2008	1558	1.5	0.1	1.4	3.3	1	0.3	2.4
2009	2047	0.8	0.3	0.5	1.1	1.5	0.4	1
2010	1789	0.7	0.4	0.3	3.9	0.4	1.2	3.9
2011	1314	4.3	0.3	4	3.1	2.6	0.8	4.8
2012	991	0.5	0	0.5	0.8	1.2	0.8	0.7
2013	651	1.4	0.5	0.9	1.2	0.8	1.7	0.8
2014	806	20.5	1.9	18.5	19.2	10.5	4.6	18.7
2015	1020	15.9	0.1	15.8	6.5	1.9	4.3	15.7
2016	539	28.2	0.9	27.3	11.3	4.8	9.5	26.9
2017	464	11.4	0	11.4	6.5	6.7	6.7	9.7
2018	1831	6.3	0.2	6.1	6.1	4.9	2.2	5.9
2019	2755	43.1	3.2	39.9	42.4	26.9	20.5	42.3
2020	2738	27.6	4.9	22.6	25.6	18	26.6	24.8
2021	2490	34.1	4.5	29.3	30.3	22.2	21.8	29.5
2022	2848	15.6	1	14.6	14.7	11.5	18.4	14.3
2023	2721	15.4	4.4	10.6	12	9.4	32.2	9.9
2024	2733	16.8	3.9	10.5	12.7	8.7	33.4	7.9
2025	2857	13.7	3.1	9.8	10.7	7	20.7	9
2026	2956	69.2	8.2	61.1	66.5	52.2	79.2	61.6

Filter rules, 2021 to 2026:

year	HTML messages	SA __STY_INVIS	SA LOW_CONTRAST	SA invisible buffer 20+	SA MPART_ALT_DIFF	Rspamd MANY_INVISIBLE_PARTS	Rspamd ZERO_FONT	Rspamd HIDDEN_TEXT	Nine declarations (Talos)
2021	2490	3.5	35.6	40.3	33.5	35.9	5	11.9	21.8
2022	2848	3	19.6	19.2	28.1	19.9	1.9	3.5	18.4
2023	2721	7.8	19.3	18.5	29.3	21.3	3	2.4	32.2
2024	2733	14.9	13.9	17.9	17.5	19.2	2	7.5	33.4
2025	2857	11.8	11.6	13.4	15.9	15.1	3	6.4	20.7
2026	2956	51.2	28.8	45	1.2	45.8	26.1	60.3	79.2

year	HTML messages	SA __STY_INVIS	SA LOW_CONTRAST	SA invisible buffer 20+	SA MPART_ALT_DIFF	Rspamd MANY_INVISIBLE_PARTS	Rspamd ZERO_FONT	Rspamd HIDDEN_TEXT	Nine declarations (Talos)

S4 Spam Archive, pooled over the periods of prior work

Years weighted by the estimated number of HTML messages in the archive.

Period	Estimated HTML messages	Sampled HTML messages	Hidden region 20+	Guo et al. rule, our analyser	Guo et al. rule, released script	Residual	Leading only
2000–2024 (Guo et al.)	3,438,842	42,650	5.8%	4.9%	3.1%	5.3%	0.5%
2003–2018 (Betts et al.)	3,301,154	22,363	5.0%	4.1%	2.5%	4.6%	0.4%
2021–2026	84,988	16,605	33.7%	30.9%	23.7%	28.8%	4.5%

S5 The released script of Guo et al. on mailbox 1 (2021 to 2026)

Run unchanged (commit 73a5677): any hidden text node in 11.4% of HTML messages [8.3, 15.4]; with the message rule of their Section 4.2 applied to its output, 8.7% [6.0, 11.9]. The script reads inline styles and presentational attributes; style sheets and classes are not applied, which is why the main text uses its thresholds inside our analyser.

S6 Checks on the second mailbox, fixed before its bodies were fetched

Check	Criterion	Mailbox 1	Mailbox 2	Outcome
W1 Share of hidden-text messages with a leading region	at least 75%	90.1%	92.2% [85.3, 97.6]	holds
W2 Share of leading regions that appear in the preview (random sample, 2021 to 2026)	at least 90%	96.9%	96.8% (1,105 regions)	holds
W3 Rule on any CSS-hidden text fires at least three times as often as the residual rule, and four other structural rules fire more often than the residual rule	at least 3	10.2	17.0	holds
W4 No tag character (other than flags) and no strict-lexicon match on the hidden side	zero	0, 0	0, 1	fails
W5 Agreement of the two implementations, standard profile	at least 99.9%	99.998%	99.951%	holds
W6 `display: none` regions: innerText below 50%, tag-stripping extractors above 95%; zero font size: innerText above 90%	as stated	21.0%, 99.7%, 98.9%, 98.1%	19.5%, 99.5%, 99.8%, 100%	holds

S7 Gmail body clipping and quoted replies (post hoc)

Gmail truncates the body view of HTML above 102 KB. 10.4% of recent HTML messages exceed this size, and 17.3% of the recent messages with residual text are among them, so part of the residual text lies beyond what that client shows. No message with residual text contains quoted-reply markup.

S8 Patterns for the nine declarations counted by Cisco Talos

Case-insensitive regular expressions applied to the HTML source (presence only):

- `font_size0`: `font-size\s*:\s*0(?![\d.])(?:px|pt|em|%)?`
- `opacity0`: `(?<![-\w])opacity\s*:\s*0(?![\d.])(?:px|pt|em|%)?`

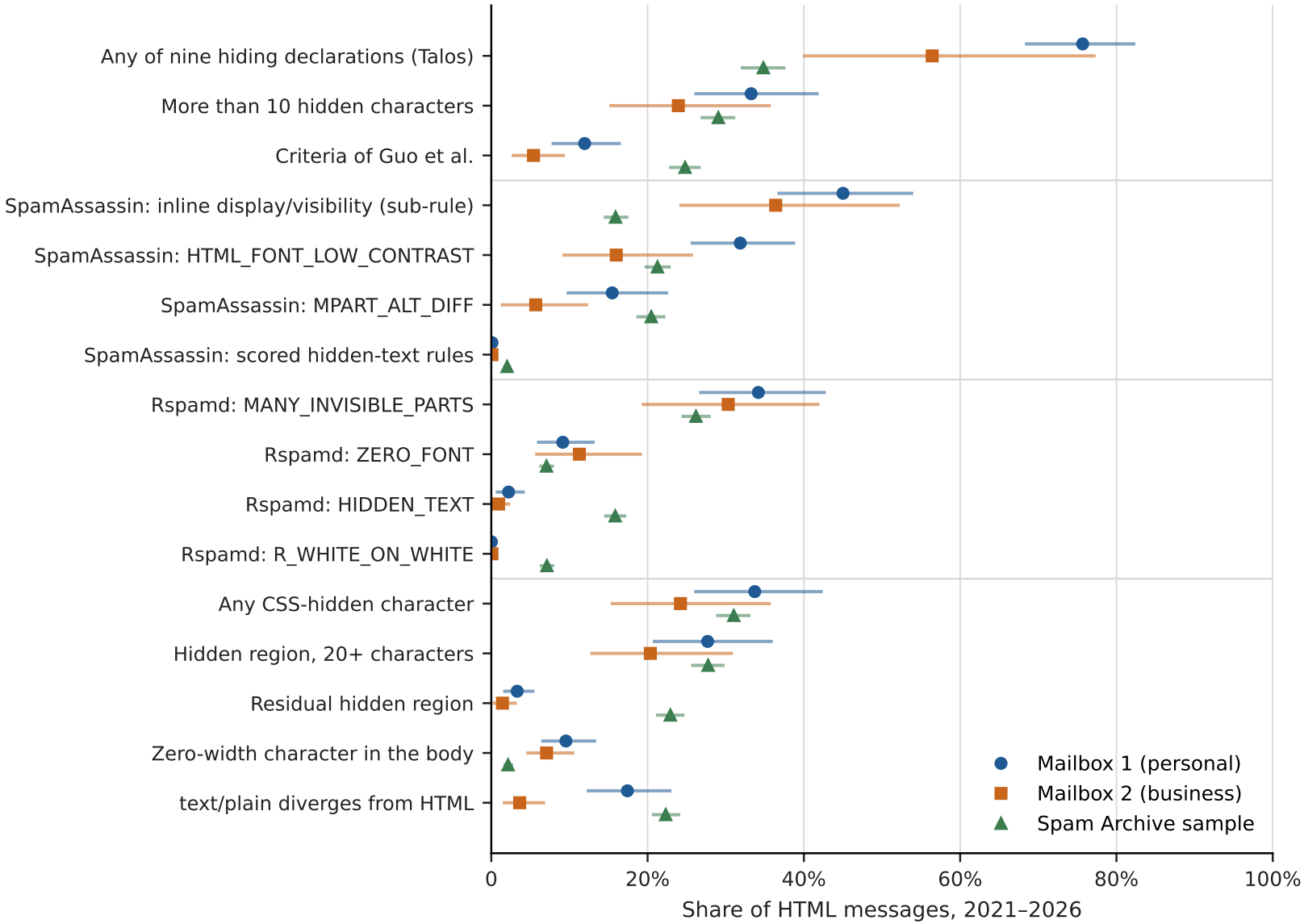


Figure S1: Share of HTML messages of 2021 to 2026 on which each criterion or rule fires: two mailboxes of ordinary mail (95% cluster-bootstrap intervals over sender domains) and the Spam Archive sample (years pooled as sampled; the bar is a sampling interval only).

- display_none: `display\s*:\s*none`
- max_width0: `max-width\s*:\s*0(?![\d.])(?:px|pt|em|%)?`
- max_height0: `max-height\s*:\s*0(?![\d.])(?:px|pt|em|%)?`
- color_transparent: `(?<![-\w])color\s*:\s*transparent`
- visibility_hidden: `visibility\s*:\s*hidden`
- width0: `(?<![-\w])width\s*:\s*0(?![\d.])(?:px|pt|em|%)?`
- height0: `(?<![-\w])height\s*:\s*0(?![\d.])(?:px|pt|em|%)?`
- mso_hide: `mso-hide\s*:\s*all`

S9 Web-archived newsletters: usable pages by capture year, provider and source

year	activecampaign (Common Crawl)	activecampaign (Wayback)	campaignmonitor (Wayback)	ctct (Common Crawl)	ctct (Wayback)	mailchimp (Common Crawl)	mailchimp (Wayback)
2013	0	0	0	264	0	0	274
2014	0	0	0	271	0	0	271
2015	1	0	93	0	61	0	272
2016	0	3	57	156	0	0	275
2017	0	0	0	273	0	0	275
2018	0	0	0	273	0	275	0
2019	0	0	0	275	0	275	0
2020	0	0	0	273	0	275	0
2021	0	0	0	273	0	272	0
2022	0	0	0	274	0	274	0
2023	0	0	0	274	0	275	0
2024	0	0	0	272	0	0	274
2025	0	0	0	275	0	0	259
2026	0	0	0	270	0	0	238

S10 Mailchimp pages by footer copyright year: share with hidden text of 20+ characters

Copyright year	Pages	Hidden 20+
2011	44	4.5%
2012	160	1.9%
2013	338	0.6%
2014	280	1.8%
2015	343	2.9%
2016	299	5.7%
2017	223	12.6%
2018	141	42.6%
2019	141	40.4%
2020	168	51.8%
2021	146	55.5%
2022	120	55.0%
2023	65	63.1%
2024	135	58.5%
2025	159	44.0%
2026	112	62.5%

S11 Readers of hidden text (table behind Figure 2)

Share of hidden regions of 20+ characters (mailbox 1, both samples, all years) whose first 200 characters are present verbatim, after removal of whitespace and padding, on each surface or in each reader's output. The last two columns were added post hoc.

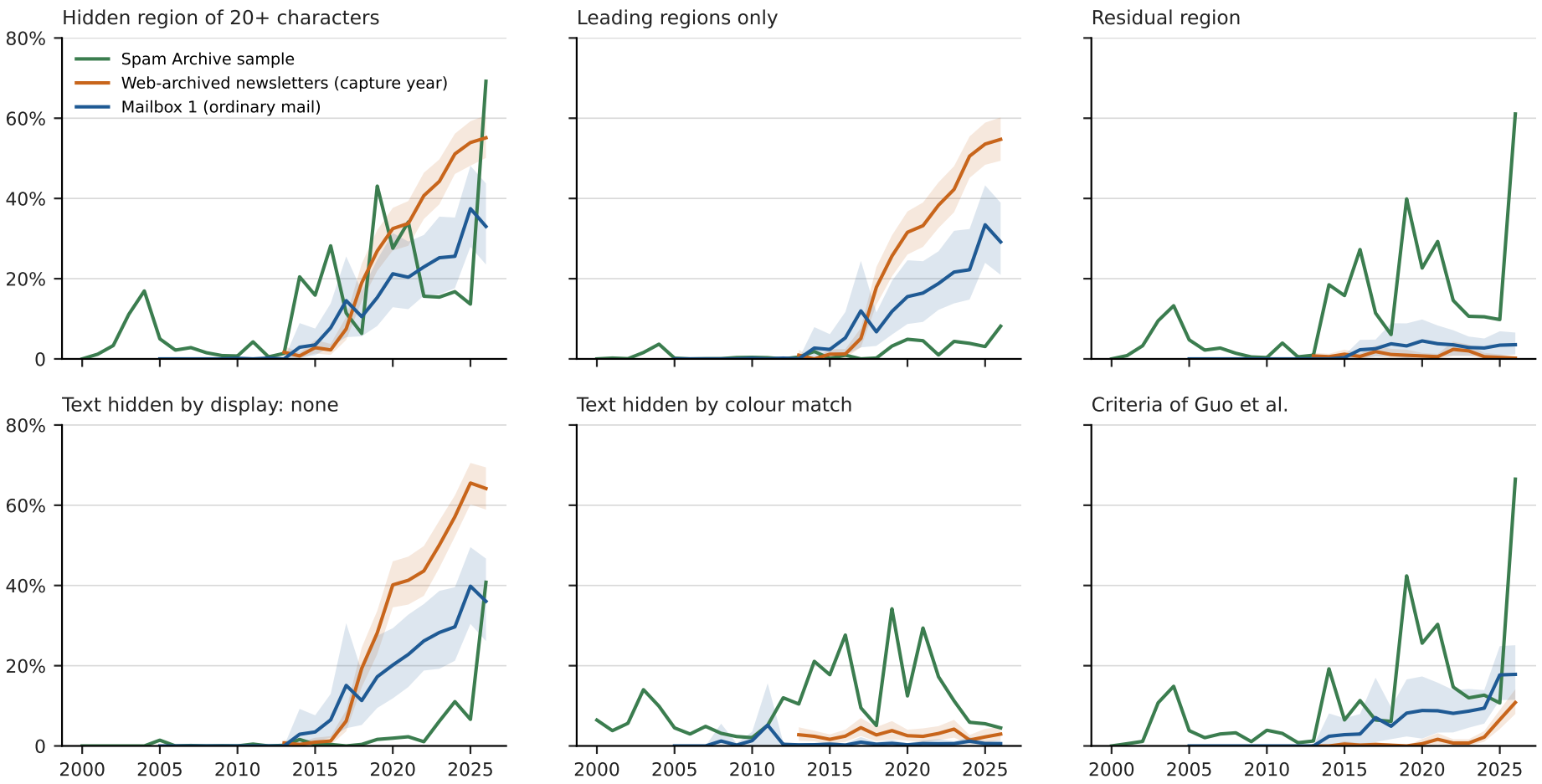


Figure S2: All six panels of the corpus comparison of Figure 3, including the hiding mechanism and the criteria of Guo et al.

Hidden regions	n	Gmail preview	Beautiful Soup	lxml	html2 text	text/plain first	Chromium innerText	Rspamd text	SpamAssassin visible text
Leading	3,876	97%	100%	99%	97%	37%	25%	18%	50%
Duplicate of visible text	413	25%	100%	100%	51%	16%	81%	72%	77%
Residual	726	20%	100%	100%	72%	33%	14%	12%	37%
`display: none`	4,543	80%	100%	99%	90%	34%	21%	23%	53%
Zero font size	240	83%	100%	100%	87%	38%	98%	3%	32%
Colour match	105	77%	100%	100%	88%	58%	99%	9%	11%
Font of 2 px or less	75	81%	100%	95%	96%	24%	100%	8%	24%
All	5,015	80%	100%	99%	90%	34%	28%	22%	51%

S12 Sensitivity of the region classes and of the clustering (post hoc)

Shares of HTML messages, 2021 to 2026, with leading regions only, a duplicate region, and a residual region, by the threshold of the duplicate class (0.8 was fixed in advance).

Corpus	n	0.60	0.70	0.80	0.90	1.0
Mailbox 1	8,732	23.9 / 2.0 / 3.2	23.9 / 1.9 / 3.3	23.9 / 1.8 / 3.3	23.9 / 1.7 / 3.4	23.8 / 1.6 / 3.5
Mailbox 2	5,891	18.6 / 1.3 / 1.1	18.6 / 1.0 / 1.4	18.5 / 1.0 / 1.4	18.5 / 1.0 / 1.4	18.4 / 0.8 / 1.7
Spam Archive sample	16,605	4.2 / 0.9 / 22.8	4.2 / 0.8 / 22.9	4.2 / 0.8 / 22.9	4.2 / 0.8 / 23.0	4.2 / 0.7 / 23.0
Web newsletters	3,230	45.5 / 0.3 / 0.7	45.4 / 0.2 / 0.9	45.3 / 0.2 / 1.0	45.3 / 0.1 / 1.0	45.3 / 0.1 / 1.0

Cells: leading only / duplicate / residual, per cent. Leading only as in Table 4 (a leading region and no residual region).

Free-mail senders (mailbox 1, sample R, HTML messages): hidden region of 20+ characters, rate [95% interval], under three treatments of the free-mail cluster (each row is a separate bootstrap, so the first row differs slightly from the interval in the main text).

Treatment	2005–15	2016–26	2021–26
merged (as specified)	1.0% [0.2, 2.4]	22.5% [17.4, 28.5]	27.7% [21.0, 36.1]
excluded	1.0% [0.3, 2.5]	22.5% [17.2, 29.2]	27.7% [20.8, 36.2]
one cluster per message	1.0% [0.3, 2.5]	22.5% [17.2, 29.0]	27.7% [20.9, 35.7]