Q8 — Extended PageRank (15 points)

Graph Structure

```
a.gov → b.mil, d.com
b.mil → d.com, e.com
c.edu → d.com
d.com → f.spam
e.com → g.spam
f.spam → g.spam
g.spam → c.edu
h.spam → g.spam
i.spam → g.spam
```

Trusted seeds: a.gov, b.mil, c.edu (trust = 1 each).

Part 1 (7 points) — Maximum β for Correct Spam Detection

Trust threshold: 0.18. All spam nodes (f, g, h, i) must have trust < 0.18.

Trust Propagation

A node with trust t and out-degree d sends $\beta \cdot t/d$ to each out-neighbor.

Node	Sources of trust	Trust value
d.com	a.gov ($\beta/2$) + b.mil ($\beta/2$) + c.edu (β)	2β
e.com	b.mil ($\beta/2$)	$\beta/2$
f.spam	d.com ($\beta \cdot 2\beta$)	$2\beta^2$
g.spam	e.com ($\beta \cdot \beta/2$) + f.spam ($\beta \cdot 2\beta^2$)	$\beta^2/2 + 2\beta^3$
h.spam	(no trusted path)	0
i.spam	(no trusted path)	0

Constraints (trust < 0.18 for all spam nodes)

- **h.spam, i.spam:** trust = 0 < 0.18. ✓ Always satisfied.
- **f.spam:** $2\beta^2 < 0.18 \;\Rightarrow\; \beta^2 < 0.09 \;\Rightarrow\; \beta < 0.3$
- **g.spam** at $\beta = 0.3$: $0.09/2 + 2(0.027) = 0.045 + 0.054 = 0.099 < 0.18$. ✓

The **binding constraint** is f.spam.

$$\boxed{\beta_{\max} = 0.3}$$

Part 2 (4 points) — Top-k PageRank as Seed Set

Answer: No, this approach would not work well on this graph.

g.spam has very high PageRank because it receives in-links from 4 nodes (e.com, f.spam, h.spam, i.spam). The spam nodes h.spam and i.spam form a link farm that artificially inflates g.spam's importance. If top-k PageRank pages are used as the seed set, g.spam would likely be selected and assigned trust = 1, making a spam page appear trustworthy.

Key insight: PageRank can be artificially inflated by spam link farms, so high PageRank does not imply trustworthiness. This is precisely why TrustRank uses known-trusted domains (.gov, .edu, .mil) rather than relying on PageRank-based seed selection.

Part 3 (4 points) — Low Early Stopping Threshold in Pixie

A low minimum visit threshold causes the random walks to **terminate prematurely**, before the visit count distribution has stabilized.

Consequence: The recommendations become **unstable and inconsistent**. Running the algorithm multiple times on the same input could produce completely different results, because the visit counts are too small to reliably distinguish between truly popular pins and pins that received visits by chance.

Tradeoff: Lower threshold → faster execution but noisier, less reproducible recommendations. The visit counts haven't converged, so the "top-1000 most visited" list is dominated by random sampling variance rather than genuine relevance signal.

Analogy from the hint: On a dense bipartite graph with >1000 nodes, stopping when the 100th most visited pin has only 2 visits means most pins have 0–2 visits — the ranking is essentially random noise.