

Zajem in analiza metapodatkov v omrežju BitTorrent s pomočjo porazdeljene razpršene tabele

Anton Luka Šijanec
izr. prof. dr. Mojca Ciglarič

Fakulteta za računalništvo in informatiko
Fakulteta za matematiko in fiziko
Univerza v Ljubljani

2026

Uvod: Motivacija

- ▶ BitTorrent (2001): protokol vsak z vsakim za izmenjavo datotek, znan po svoji uporabi za kršenje avtorskih pravic.
- ▶ Zanima nas, za kaj je uporabljen (legalnost in kategorije vsebine), prebivalci katerih držav predstavljajo največji delež uporabnikov in druge tehnične podrobnosti.
- ▶ Problem: Kako zajeti pošten vzorec prenesene vsebine?

Uvod: Motivacija

- ▶ BitTorrent (2001): protokol vsak z vsakim za izmenjavo datotek, znan po svoji uporabi za kršenje avtorskih pravic.
- ▶ Zanima nas, za kaj je uporabljen (legalnost in kategorije vsebine), prebivalci katerih držav predstavljajo največji delež uporabnikov in druge tehnične podrobnosti.
- ▶ Problem: Kako zajeti pošten vzorec prenesene vsebine?
 - ▶ Ga zajemamo iz sledilniških strežnikov, ki koordinirajo uporabnike?
 - ▶ Ga prestrezamo na usmerjevalnikih omrežnih operaterjev?

Uvod: Motivacija

- ▶ BitTorrent (2001): protokol vsak z vsakim za izmenjavo datotek, znan po svoji uporabi za kršenje avtorskih pravic.
- ▶ Zanima nas, za kaj je uporabljen (legalnost in kategorije vsebine), prebivalci katerih držav predstavljajo največji delež uporabnikov in druge tehnične podrobnosti.
- ▶ Problem: Kako zajeti pošten vzorec prenesene vsebine?
 - ▶ Ga zajemamo iz sledilniških strežnikov, ki koordinirajo uporabnike?
 - ▶ Ne, nimamo dostopa.
 - ▶ Ga prestrezamo na usmerjevalnikih omrežnih operaterjev?
 - ▶ Ne, nimamo dostopa. Pa tudi tak vzorec bi bil geografsko pristranski.

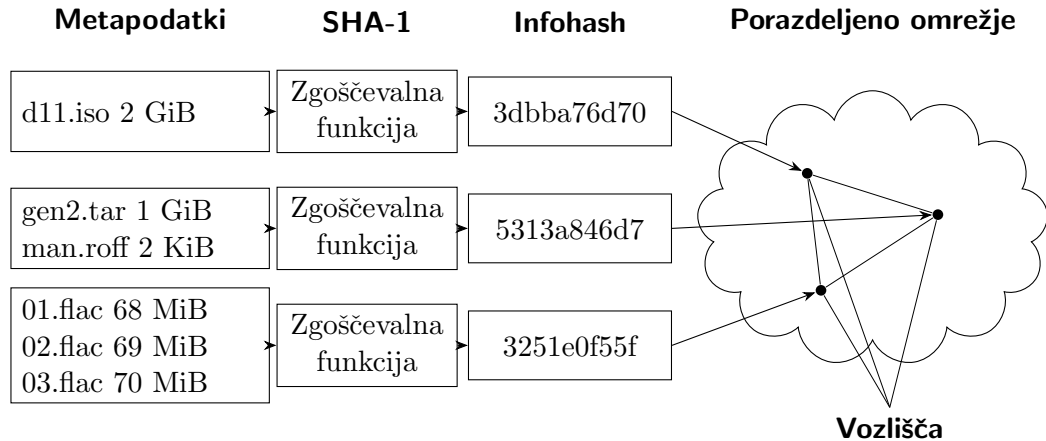
Uvod: Cilji

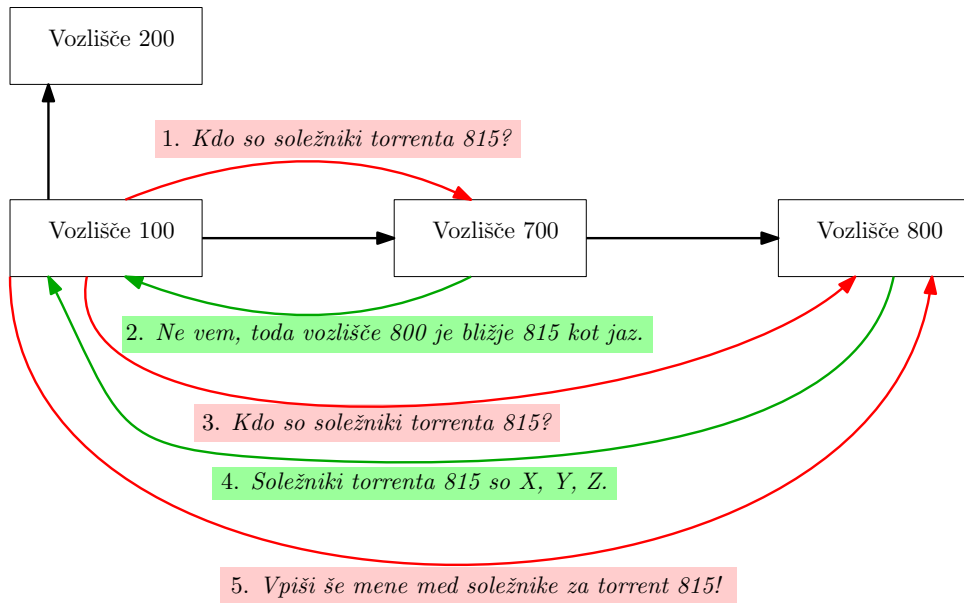
- ▶ Leta 2008 standardizirajo uporabo DHT za iskanje soležnikov. BitTorrent ne potrebuje več centralnih strežnikov. Podatki so shranjeni porazdeljeno pri uporabnikih omrežja.
- ▶ Kot uporabnik omrežja imamo shranjen majhen del podatkov.
- ▶ Cilj: izdelati program, ki beleži poizvedbe po soležnikih torrentov, ki jih prejema sodelovanjem v omrežju DHT. Te poizvedbe vsebujejo identifikator torrenta, s katerim lahko pridobimo metapodatke o torrentu (imena datotek, njihove velikosti in naslove IP prenašalcev).
- ▶ Cilj: Pridobljene metapodatke analizirati in sklepati o lastnostih omrežja BitTorrent.

Uvod: za nalogo potrebne osnovne obeh protokolov

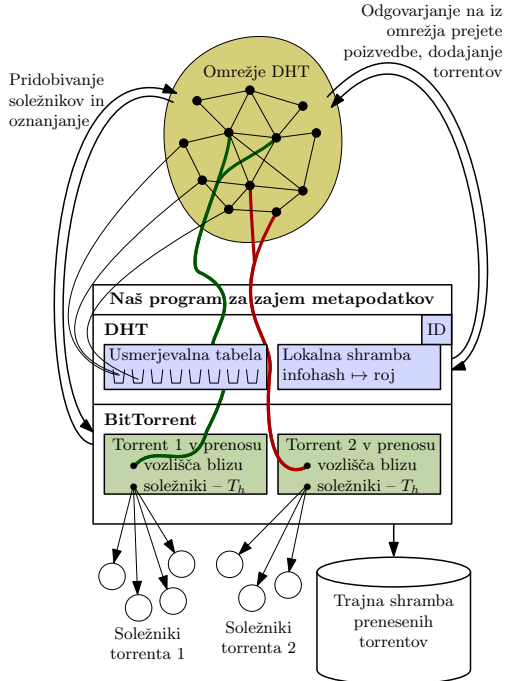
- ▶ BitTorrent (TCP): Omogoča prenos metapodatkov od odjemalca za podan *infohash*. Vsebine datotek **nismo** prenašali (bilo bi ilegalno).

Porazdeljena razpršena tabela (DHT) je na visokem nivoju slovar *infohash* $\mapsto$ soležniki. Vozlišča (računalniki) hranijo vrednosti pod tistim ključem, ki je blizu* fiksnega naključnega ID. Poizvedbo se po UDP vozliščem pošilja iterativno, na vsakem koraku smo bodisi preusmerjeni na *infohashu* bližje vozlišče, dokler ne najdemo vozlišča, ki hranijo iskan ključ.





Iterativna poizvedba v povprečju obišče $\log_2 V(G)$ vozlišč.

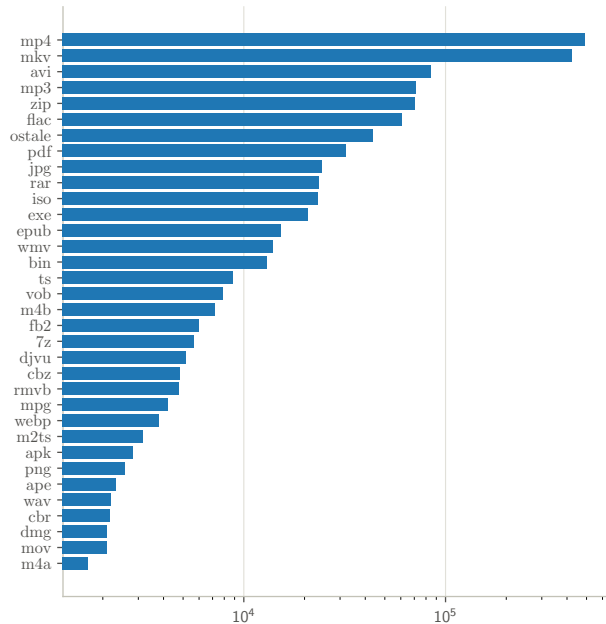


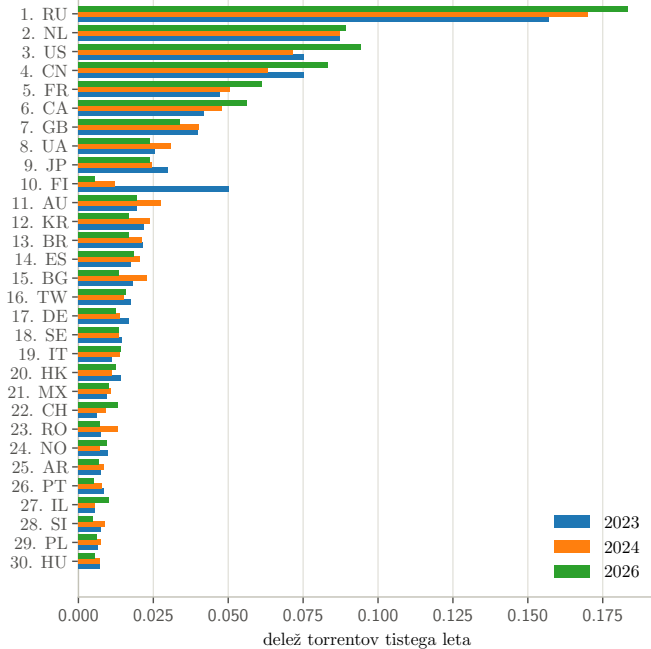
Metodologija: Program za zajem

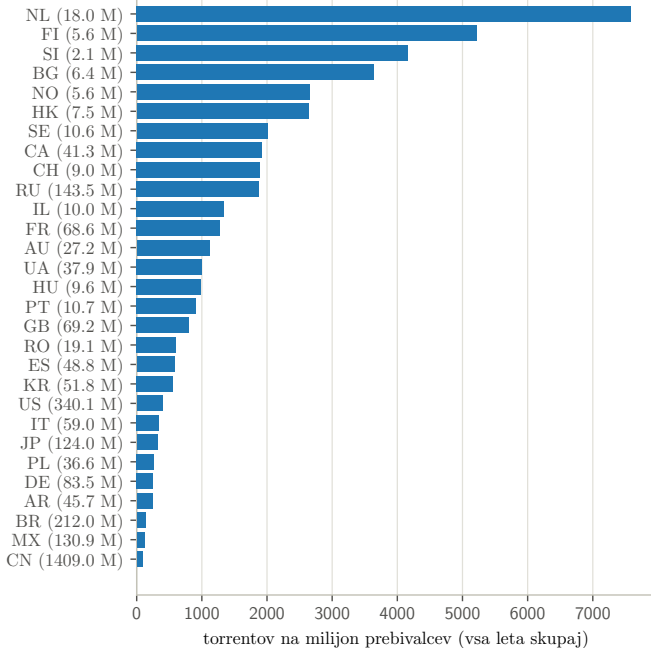
- ▶ Spisali smo odjemalec v jeziku c, ki sodeluje kot vozlišče v DHT.
- ▶ Zabeleži si *infohash* vsake poizvedbe, ki mu je poslana (in nanjo po najboljših močeh odgovori).
- ▶ Prenese metapodatke o torrentu vsakega tako najdenega *infohasha*.
- ▶ Obstoječe knjižnice za BitTorrent+DHT niso narejene za tako beleženje poizvedb. Predelave smo raje implementirali sklad BitTorrent+DHT ročno.

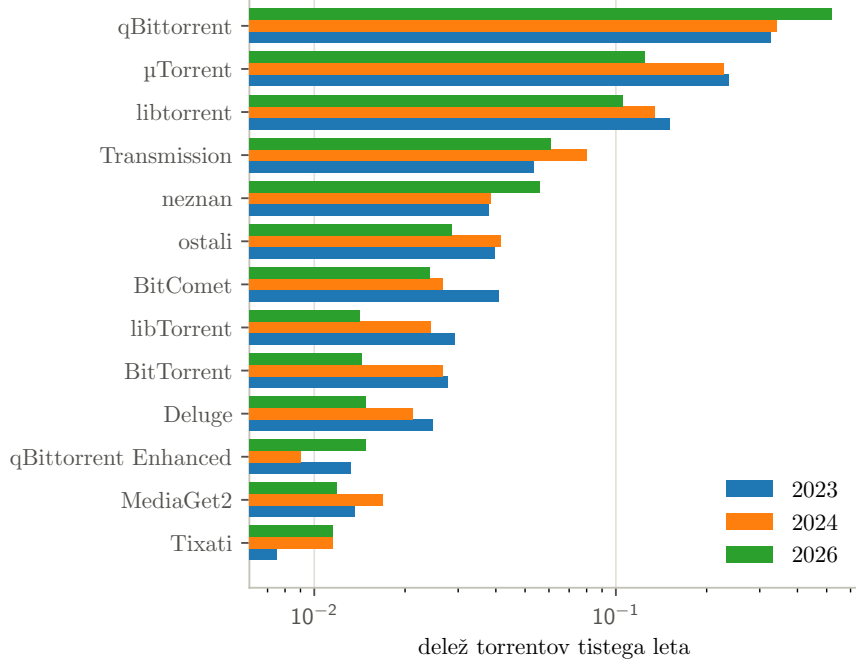
Rezultati: pregled in vsebina

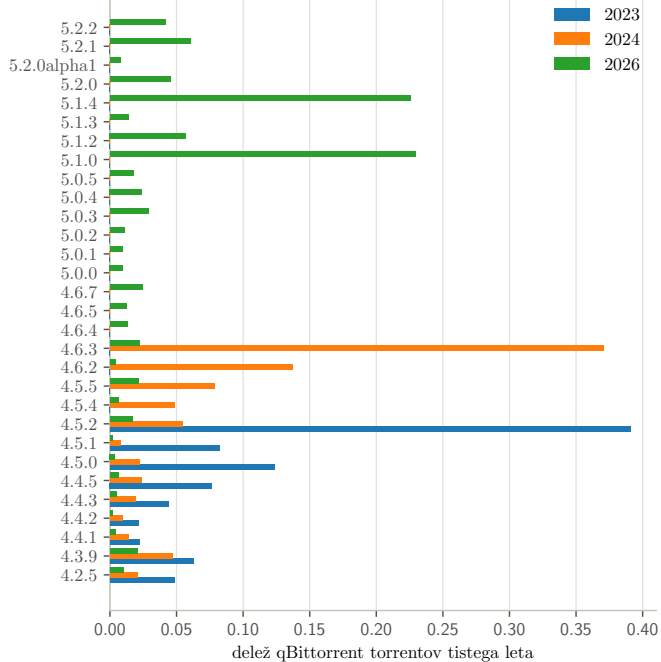
- ▶ Torrente smo na predstavljen način zbirali v letih 2023, 2024 in 2026 in skupaj zbrali 1 546 563 torrentov.
- ▶ Z jezikovnim modelom smo torrentom določili legalnost in kategorijo.
 - ▶ Piratski torrenti: 93,00 %, ilegalnih še 1,86 %
 - ▶ Legitimnih: 2,36 %, nedoločljivih 2,79 %
 - ▶ Po kategorijah: 28,9 % pornografije, 15,2 % TV oddaj in serij, 14,7 % filmov, 11,5 % glasbe
- ▶ Imamo samo imena datotek, ne pa vsebine! Prenos vsebine bi zahteval 9,112 PiB prostora.
- ▶ Največ uporabnikov torrentov je iz Rusije, Nizozemske, ZDA, Kitajske, Francije, ...



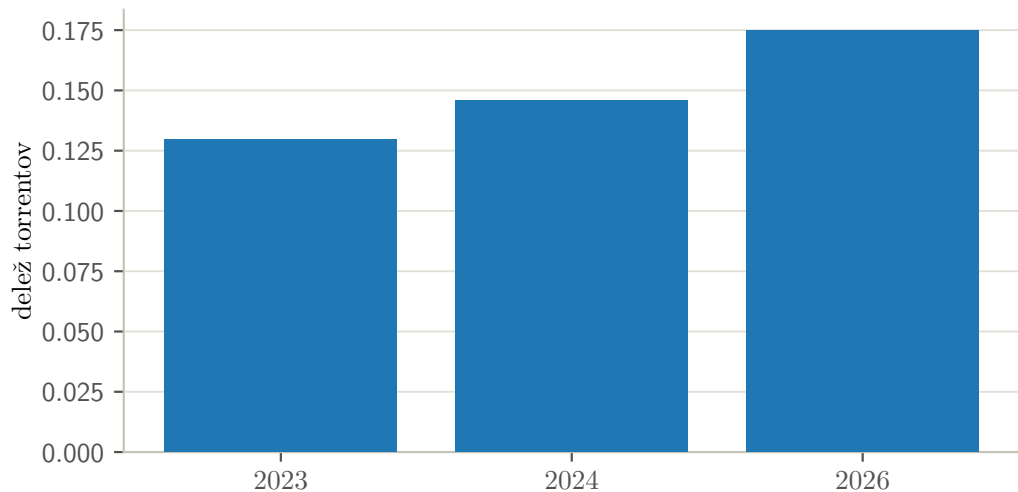


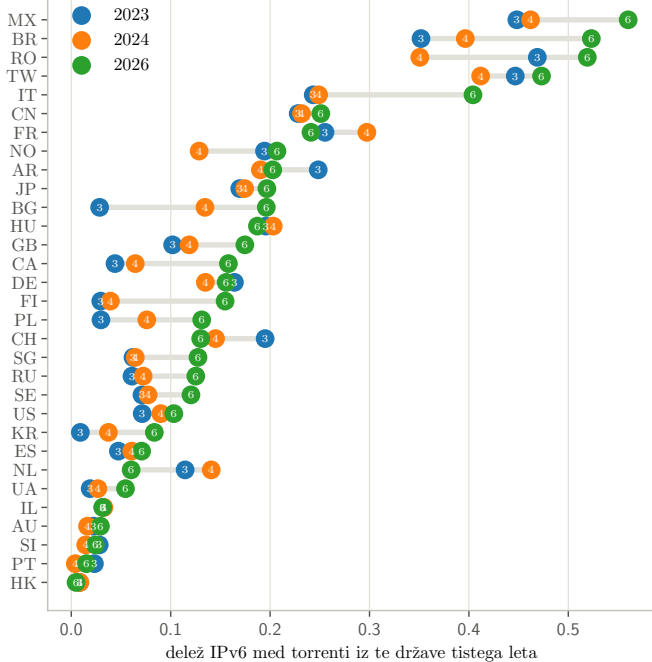






Rezultati: stranski produkt – razširjenost IPv6





Zaključek

- ▶ Naredili smo program za zajem, zajemali metapodatke v treh letih in prikazali osnovne lastnosti omrežja BitTorrent.
- ▶ Potrdili smo domnevo, da je BitTorrent v glavnem uporabljen za piratsko vsebino.
- ▶ Ideje za prihodnje raziskave: opazovanje celotnega roja torrenta (skozi čas), stabilnejši in neprestan zajem, prenos vsebine datotek za določanje uporabljenih kodekov in vsebine stisnjenih arhivov.

Hvala za pozornost!



Dodatek: Ocenjevanje pravilnosti klasifikacije (Cohenova κ)

- ▶ Ročno smo kategorije določili le 119 torrentom.
- ▶ `sklearn.metrics.cohen_kappa_score(llm_kategorije, moje_kategorije)`
- ▶ $\kappa \in [-1, 1]$, kjer je 0 naključna izbira in 1 popolno ujemanje
- ▶ $\kappa = 0,738$ za kategorije
- ▶ Večinski klasifikator, ki vedno izbere pornografijo, bi imel $\kappa = 0.294$.
- ▶ Lahko primerjamo tudi po robnih porazdelitvah, recimo: Ročno smo $\frac{113}{119} = 94,95\%$ torrentov označili za piratske, jezikovni model pa je za piratske označil 93,00 % torrentov (na vseh 2799).



Top 50 countries for total BitTorrent downloads during first half of 2012 per capita:

	Country Name	Total downloads	Approx Songs+	Population
1	Australia	19,232,252	154,242,661	21,766,711
2	Ireland	3,434,737	27,546,591	4,670,976
3	Slovenia	1,416,433	11,359,793	2,000,092
4	Canada	23,959,924	192,158,590	34,030,586
5	United Kingdom	43,263,582	346,973,928	62,698,362
6	Norway	2,827,508	22,676,614	4,691,849
7	Italy	33,158,943	265,934,723	61,016,804
8	Portugal	5,607,910	44,975,438	10,760,305
9	Croatia	2,305,095	18,486,862	4,483,804
10	Greece	4,933,478	39,566,494	10,760,136
11	Sweden	4,083,279	32,747,898	9,088,728
12		1,252,222	15,156,222	1,742,727

Dodatek: Zajem na raziskovalni infrastrukturi FRI ni bil neprekinjen

