b-tagging techniques

prospects at the LHC experiments

Clemens Lange

Workshop on single top physics and fourth generation quarks

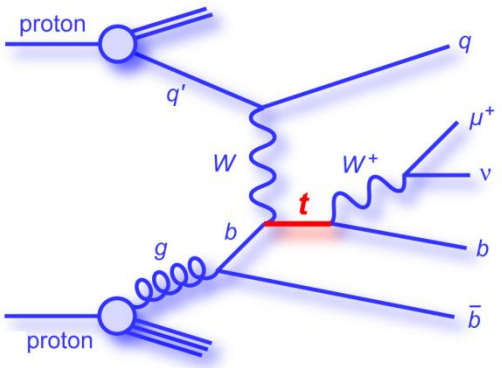
Hamburg, 15/09/09

HELMHOLTZ ASSOCIATION

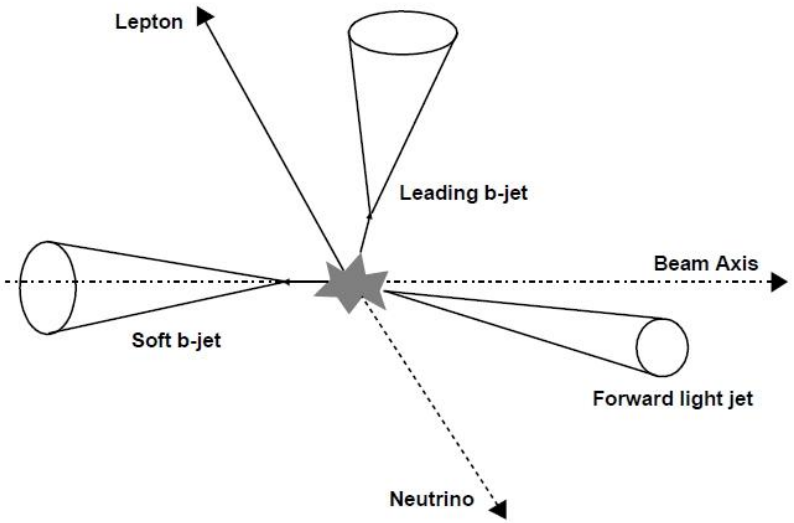
DESY

reminder: single top event selection

- selection for both CMS and ATLAS similar
- follows event topology:
  - lepton trigger (muon or electron, p_T > 15-20 GeV)
  - exactly one well-identified lepton (20-25 GeV)
  - some missing E_T requirement
  - 2-4 jets, p_T > 30 GeV
  - b-tagged jet
- what's so special about the b-tag?



single top t-channel Feynman diagram [9]



single top t-channel event topology [2]

What do we need b-tagging for?

- expected number of events selected in 200 pb^{-1} after application of the preselection cuts:

Process	no b-tag	b-tag
t-channel (signal)	**2031 ± 22**	**940 ± 16**
Wt-channel	810 ± 13	347 ± 9
$t\bar{t}$ (l+jets)	12620 ± 45	6983 ± 36
W (→lν)+jets	81813 ± 179	2678 ± 32
Z (→ll)+jets/diboson	4672 ± 32	206 ± 7
Total	**101946 ± 189**	**11156 ± 52**
S/B	**0.02**	**0.09**

Keep in mind for later 50% efficiency loss.

[1]

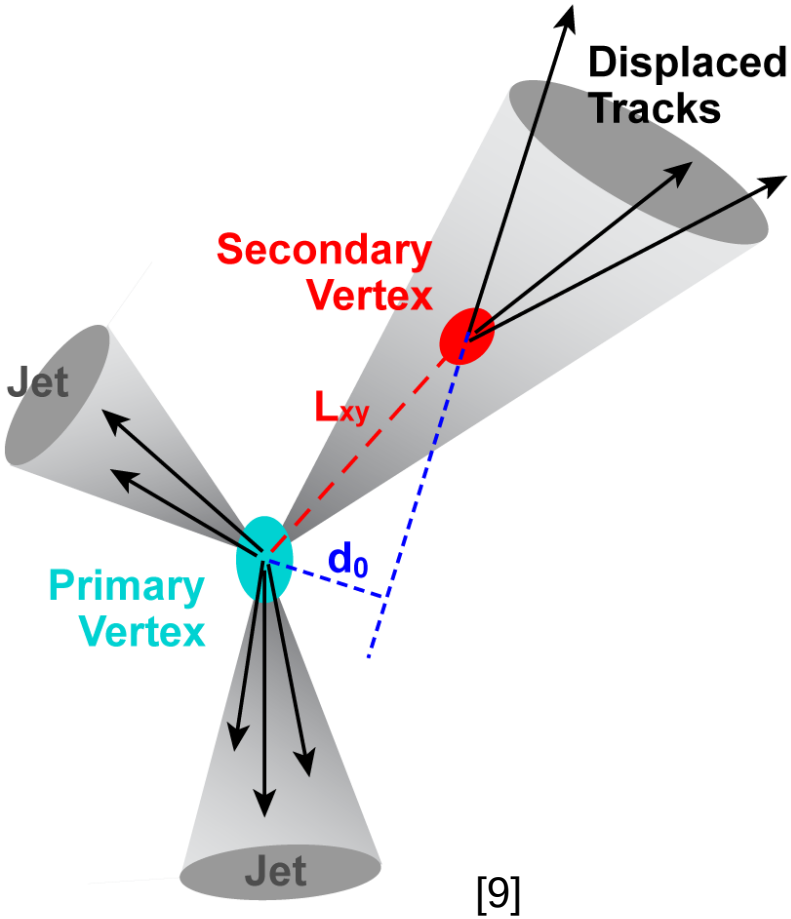
- single top signal disappears in background
- requiring b-tag improves this situation considerably!
- but what is b-tagging?

the plan for this talk

- what is b-tagging?
- ingredients to b-tagging
- issues of b-tagging
- b-tagging techniques and expected performance
  - simple secondary vertex (SV0)
  - track counting
  - JetProb
  - soft muon tagging
  - multi-dimensional likelihood taggers
- data-taking
- sum-up

what is b-tagging?

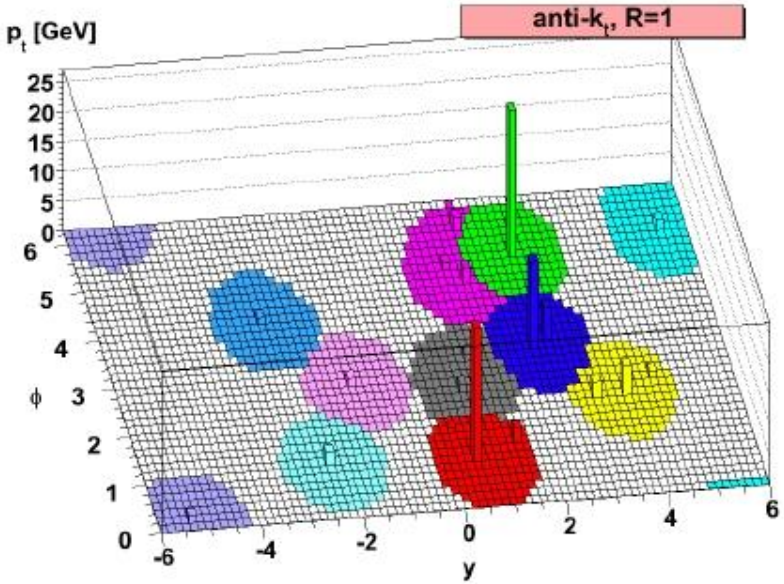
- identification of b-hadrons (containing b-quarks) in jets
- why can this be done?
- consider properties of jets containing b-hadrons to distinguish from jets containing only lighter quarks:
  - fragmentation is hard, b-hadron retains about 70% of the original b quark momentum
  - relatively high mass of b quark (> 5 GeV) → decay products may have larger transverse momentum with respect to jet axis, opening angle
  - long lifetime of b-hadrons of ~1.5 ps ($c\tau \approx 450$ μm) → 3 mm for 50 GeV hadron

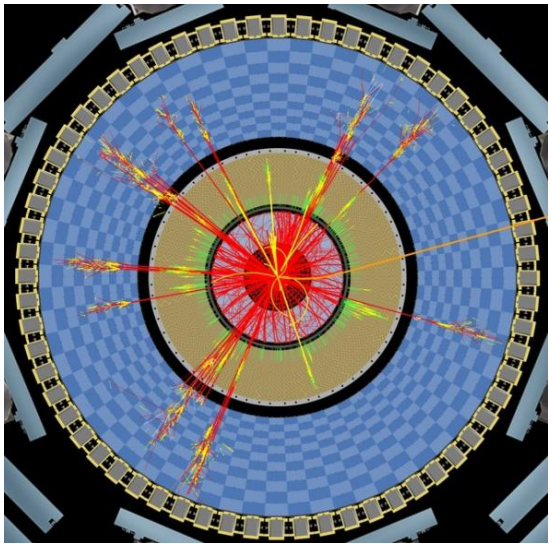


[9]

ingredients to b-tagging

- jets are built from calorimeter entries
- b-tagging relies on tracking in the inner detector
- can only use well-identified tracks:
  - sufficient pixel and strip hits
  - minimum transverse momentum
  - good track fit
- **try to find displaced tracks and/or secondary vertex**
- additionally: soft leptons in the jet

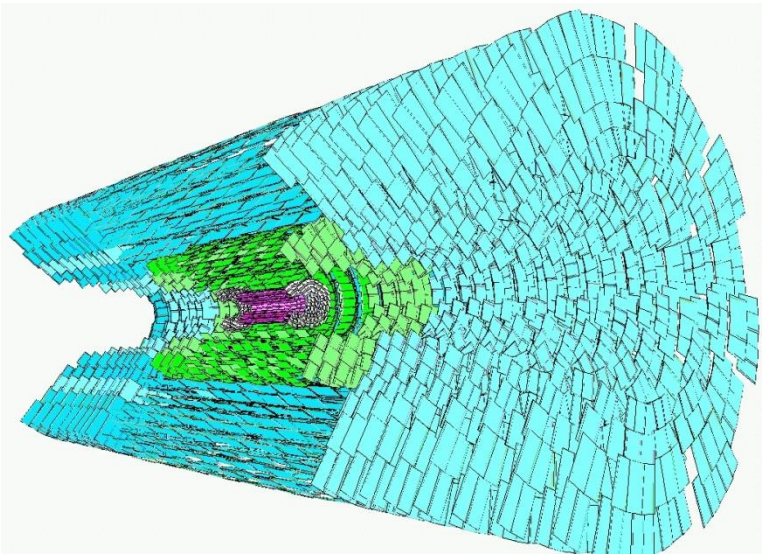


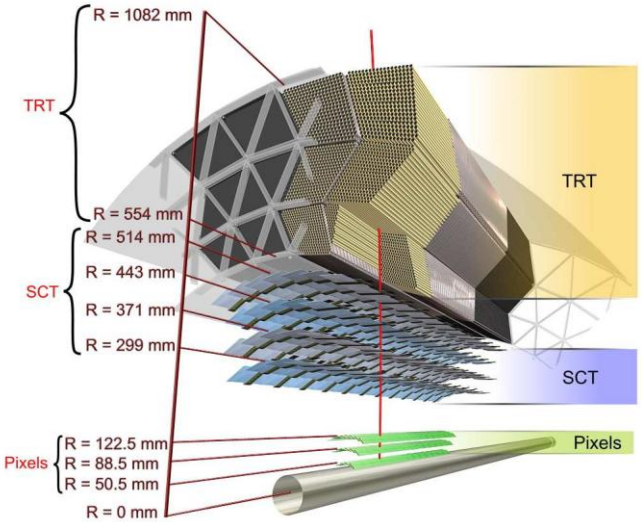


ingredients to b-tagging

- track reconstruction depends on detector resolution, especially pixel detector:

Pixel Detector	CMS	ATLAS
resolution ($r\phi$ x z)	15-20μm both	10μm x 115μm
pixel size	100 x 150μm^2	50 x 400μm^2
Magnetic field	4T	2T





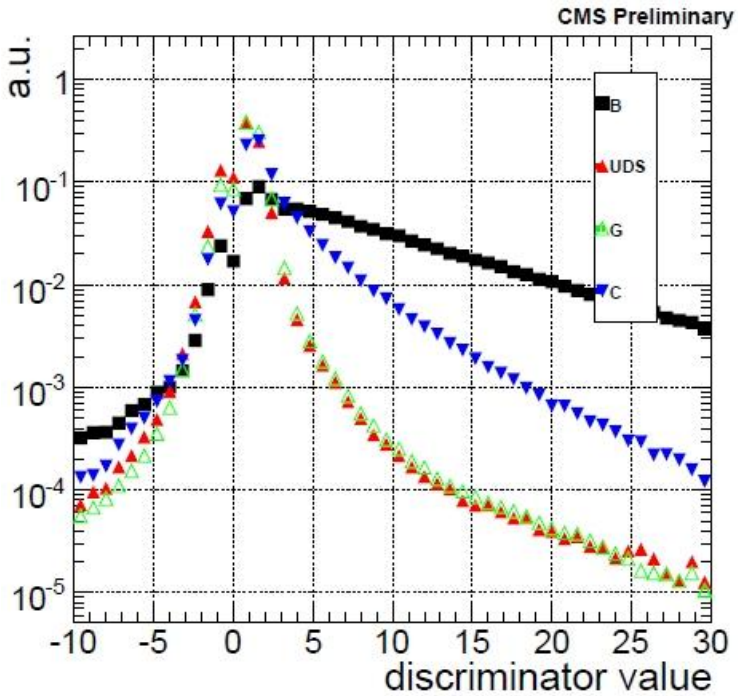
issues of b-tagging

- if b-tagging was easy, there wouldn‘t be so many different algorithms
- misalignment spoils track reconstruction and resolution
- need a well-understood detector to correct for misalignment
- further performance degradation:
  - tracker material
  - dead pixels
  - Monte Carlo simulation
- c hadron decays look similar
- some dependence on jet algorithms
- usually minimum jet transverse energy 15 GeV, minimum track transverse momentum 0.5-1 GeV

- fair enough, how do these algorithms work and how do I use them?

working with b-tagging

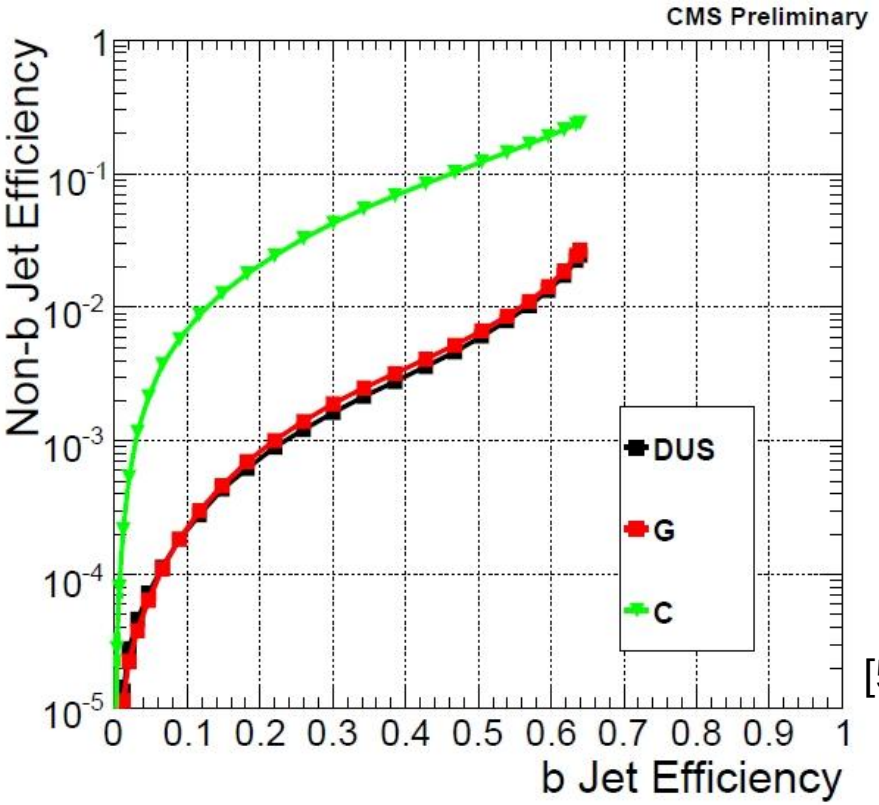
- each algorithm has a discriminator value (single number)
- cut on this value to select region in efficiency vs. purity phase space
- **increasing b jet efficiency goes along with decreasing light jet rejection**
- let's have a look at the actual algorithms, from simple and robust first data taggers to advanced ones



[5]

simple secondary vertex

- reconstructs two-track vertices significantly displaced from primary vertex
- removes K^0_s, Λ^0 conversions and material interactions
- fits inclusive vertex from tracks in all good two-track vertices
- removes the track with largest X^2 contribution until fit prob. of vertex acceptable
- tries to reincorporate initially failing tracks
- discriminator: signed decay length significance



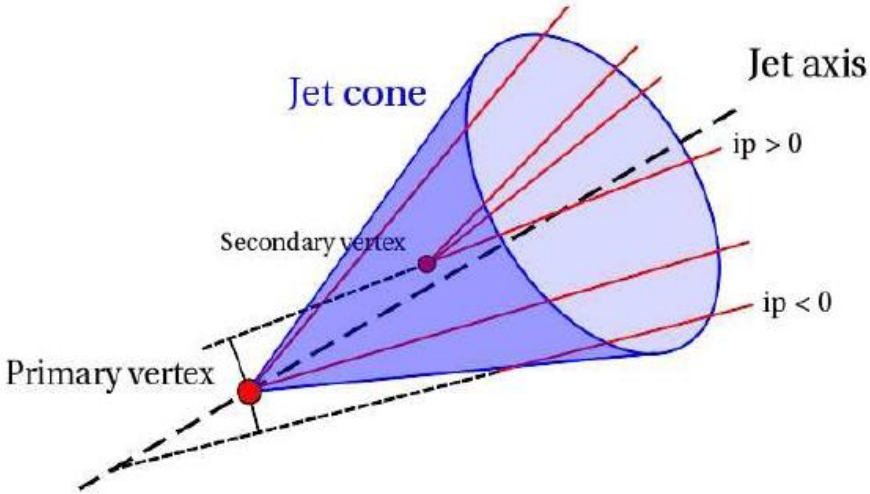
[5]

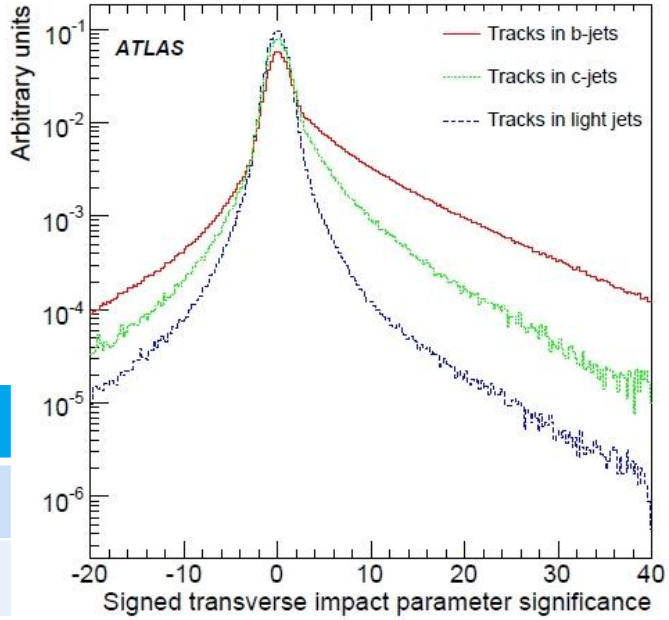
CMS study	uds rejection
50% efficiency	~200
60% efficiency	~65

most simple → first data tagger

track counting

- count tracks in a jet that have a large (transverse) impact parameter d_0
- sign: positive if decay occurs downstream the jet
  - around zero for light jets, positive for b decays
- calculate impact parameter significance d_0/σ_{d_0}
- sort by significance, cut on 2nd or 3rd track significance





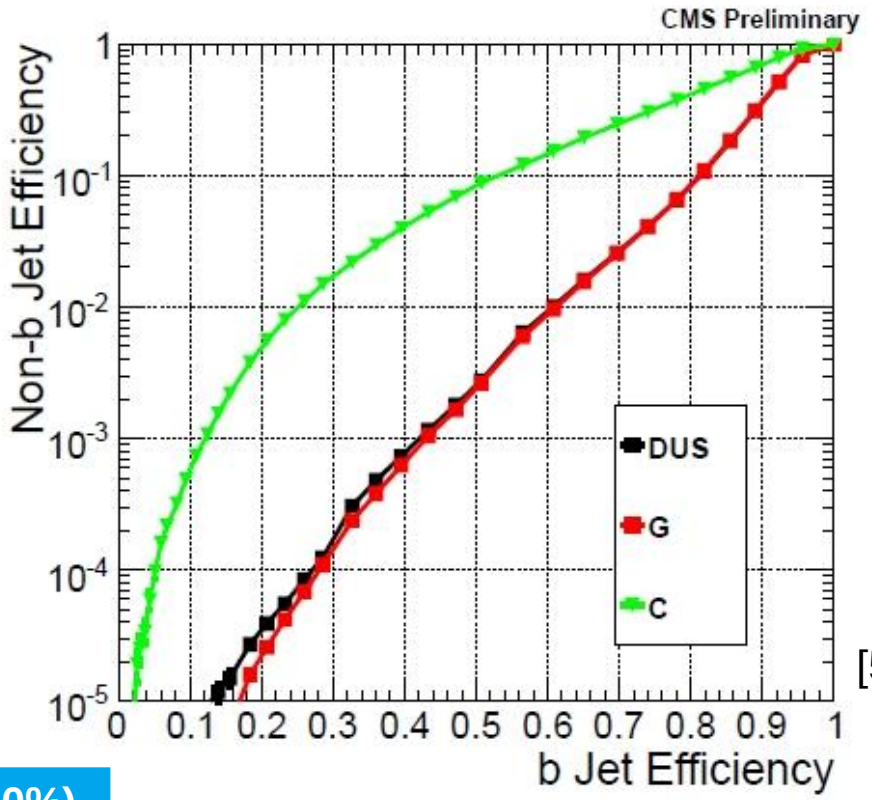
[1]

CMS study	uds rej (50%)	uds rej (60%)
high purity	~500	~65
high efficiency	~350	~125

simple and robust → first data tagger

jet probability

- combine impact parameter of all tracks in jet:
- compare each track to resolution function & determine probability that track originates from primary vertex
- affected by misalignment (PDFs have to be remade)



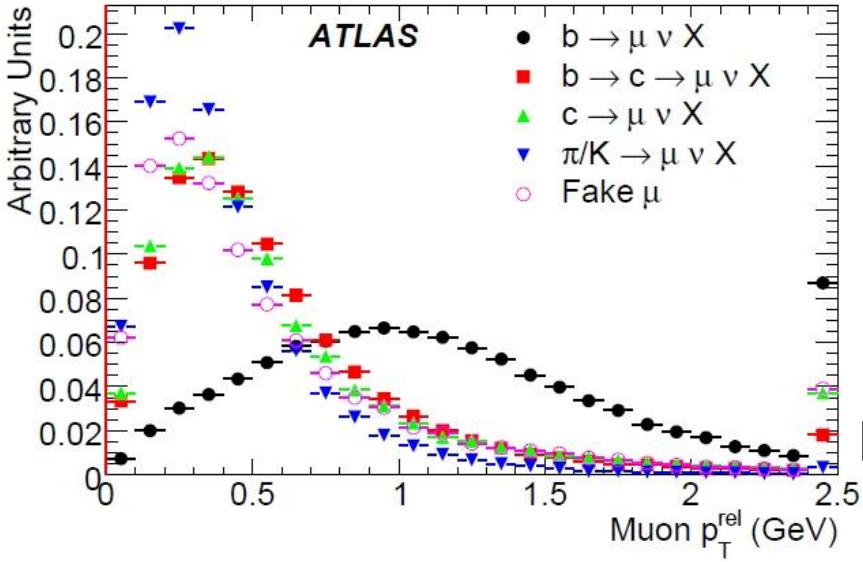
[5]

	uds rej (50%)	uds rej (60%)
ATLAS ($t\bar{t}$ evts)	~92	~31
CMS	~500	~100

soft lepton taggers

- b and c hadrons can decay leptonically (maximum 10% (directly) +11% (via c) of b jets)
- find muon associated to jet
- discrimination via 1-dimensional likelihood ratio using relative transverse momentum of muon to jet axis
- also possible for electrons (lower performance)

most likely to be used for cross-calibration



[1]

ATLAS study	uds rej
10% efficiency ($t\bar{t}$ evts)	~380

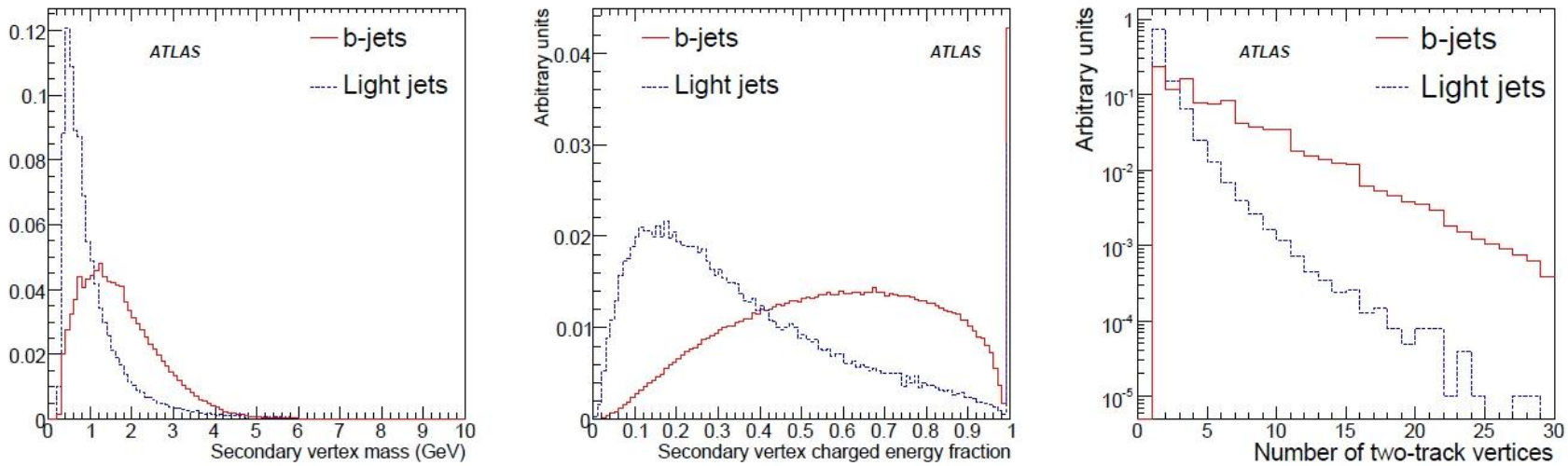
multi-dimensional likelihood taggers (1)

- CMS: combine several variables from impact parameter and secondary vertex algorithms and put into likelihood ratio, e.g.
  - vertex mass
  - energy fraction vertex/jet
  - number of two track vertices

measured value

$$W_{Jet} = \sum_{i=1}^{N_{tracks}} \ln W_i = \sum_{i=1}^{N_{tracks}} \ln \frac{b(S_i)}{u(S_i)}$$

- ATLAS: separate approach for impact parameter and secondary vertex, similar variables
- likelihood ratios from Monte Carlo distributions (b jet and light (u) jet hypothesis)

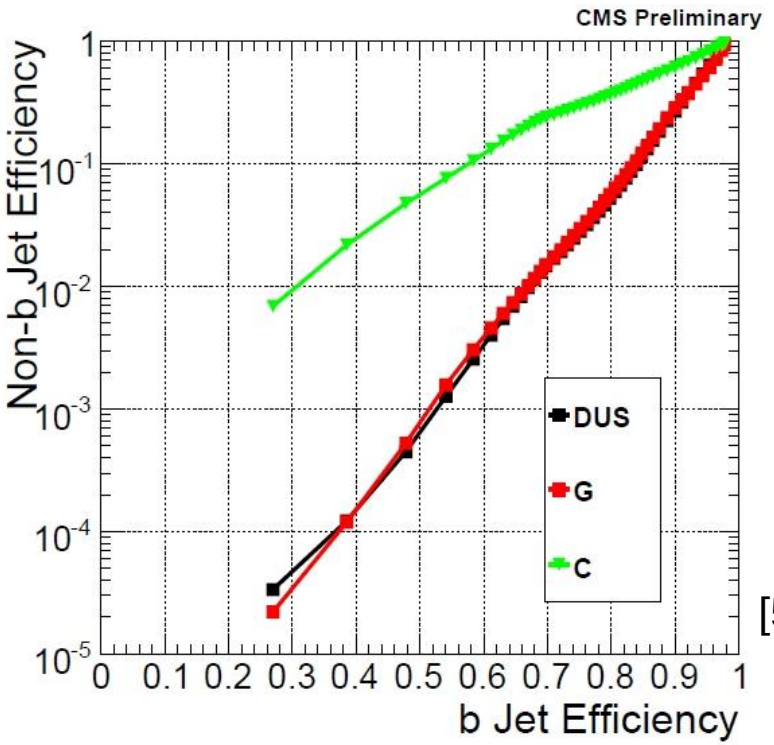


[1]

multi-dimensional likelihood taggers (2)

- CMS: combined secondary vertex
- ATLAS: IP3D+SV1
- ATLAS: new JetFitter algorithm
  - exploits topological structure of weak decays in jet in addition to other discriminators

best light jet rejection, but not suitable for first data

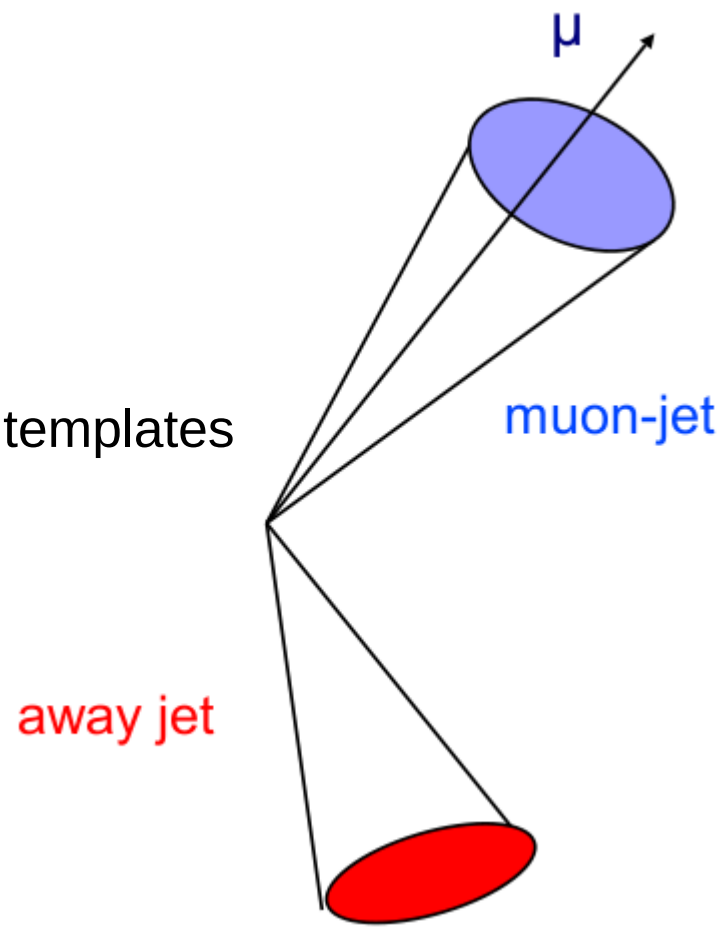


[5]

	uds rej (50%)	uds rej (60%)
ATLAS ($t\bar{t}$ evts)	~593	~180
CMS	~2000	~400

b-tagging efficiency from data

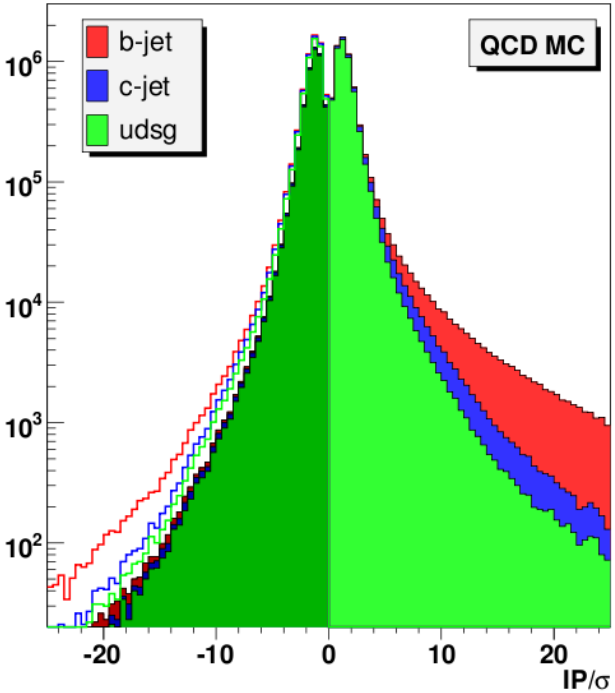
- As example: simple tag & probe method with flavour content fit
- use di-jet events
- assume b-jets are produced in pairs
- require soft muon tag in one jet
- probe the b-tag in the other jet
- Fit heavy flavour content using muon p_T^{rel} templates
- uncertainties given in backup



simple, but depends on Monte Carlo templates

light jet rejection from data

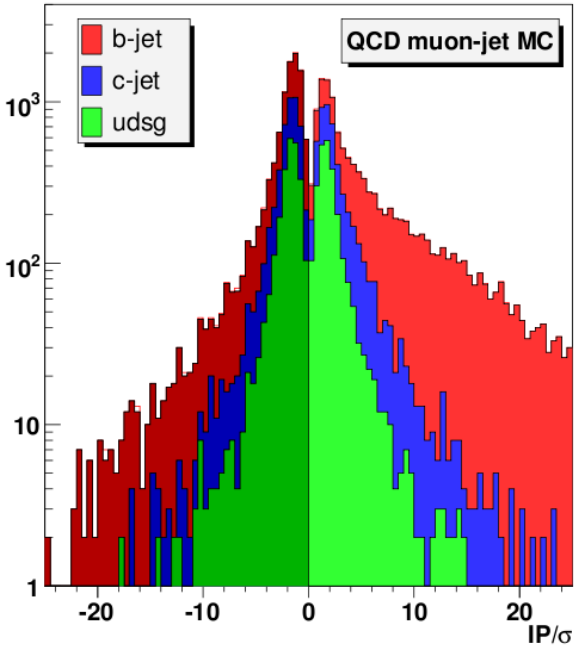
- Light flavour jet observables approximately symmetric around zero
- use sign of b-tagging observables
- mistag measurement using negative side
- Idea: flip the sign and you get a b-tagging algorithm that will behave as if the jet was always a light flavour jet



IP of all tracks

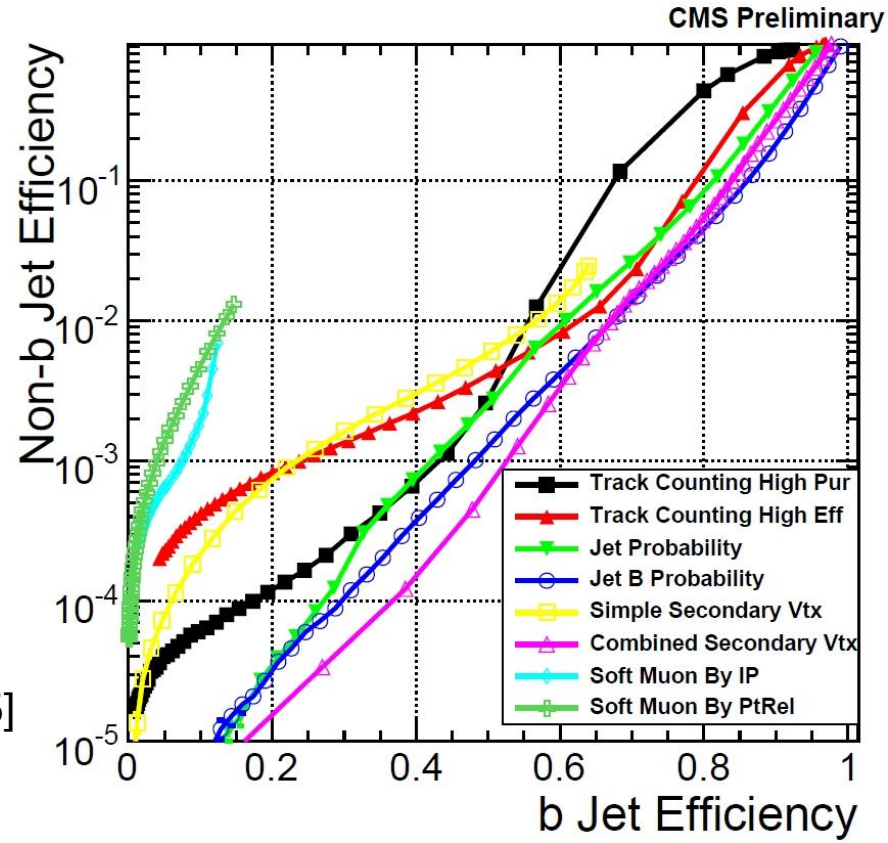
"Sign flip" effect can be estimated from b-enriched QCD *(muon in jet)*
Remaining asymmetry: "R_{light}" factor

$$\varepsilon_{data}^{mistag} = \varepsilon_{data}^{-} \cdot R_{light}$$

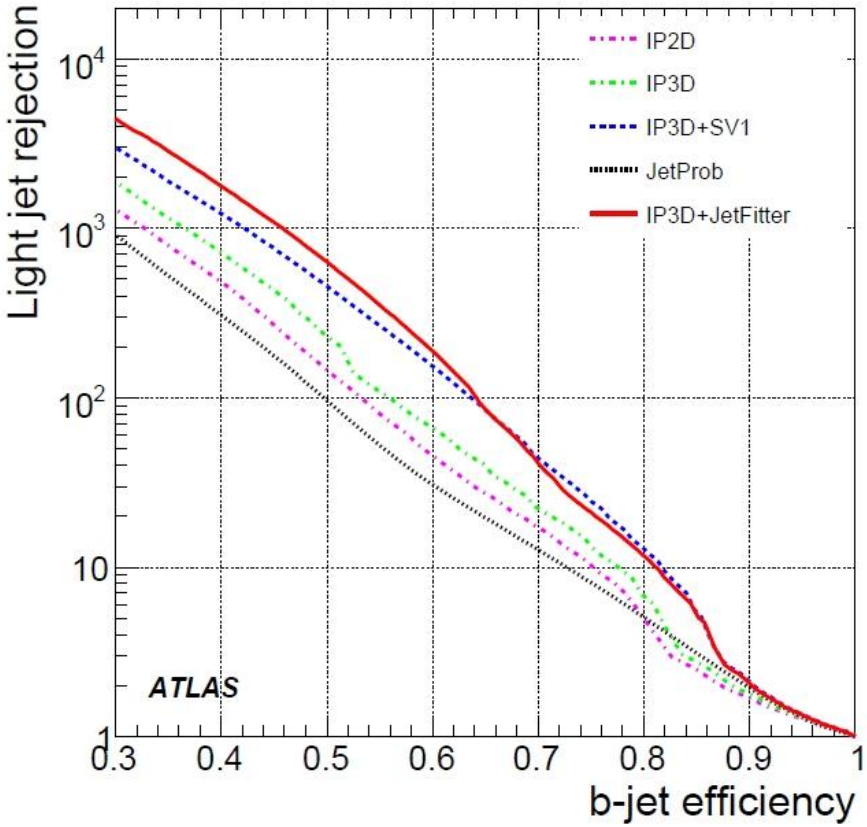


[8]

summary



[5]



[1]

- b-tagging is not easy
- several algorithms available, suitable for different purposes
- working point 50-60% efficiency, mind light jet contamination (getting better with more data)

Thank you… and here are the references

(1) Expected performance of the ATLAS Experiment, Detector, Trigger and Physics, ATLAS Collaboration, CERN-OPEN-2008-20.

(2) Mass measurements of the top quark in electroweak production channels at ATLAS, M. Barisonzi, UMI-C8-26001.

(3) Prospect for Single Top Quark Cross-Section Measurements, ATLAS Collaboration, ATL-PHYS-PUB-2009-042.

(4) Prospects for the measurement of the single-top t-channel cross section in the muon channel with 200 pb−1 at 10 TeV, CMS Collaboration, TOP-09-005.

(5) Algorithms for b Jet Identification in CMS, CMS Collaboration, BTV-09-001.

(6) Effect of misalignment on b-tagging, CMS Collaboration, BTV-07-003.

(7) b tag efficiency from System 8 & Ptrel Method, CMS Collaboration, BTV-07-001.

(8) Measuring uds mistag rate of b tag using negative tags, CMS Collaboration, BTV-07-002.

(9) http://www-d0.fnal.gov/Run2Physics/top/singletop_observation/

additional transparencies

BACKUP

Baseline track selection

ATLAS

- ≥ 7 precision hits (pixel or micro-strip), high quality: 2 pixel hits, one of them in b-layer
- $|d_0| < 2$ mm (1 mm)
- $|z_0 - z_{pv}| \sin\theta < 10$ mm (1.5 mm)
- $p_T > 1$ GeV

CMS

- ≥ 8 silicon (pixel + strip hits), 2 pixel hits
- $|d_0| < 2$ mm
- $|z_0| < 170$ mm
- X^2/ndf of track fit < 5.0
- $p_T > 1$ GeV

ATLAS jet probability

Another approach is to combine the impact parameter of all the tracks in the jet. JetProb is an implementation of the ALEPH tagging algorithm [14], used extensively at LEP and later at the Tevatron. The signed impact parameter significance d_0/σ_{d_0} of each selected track in the jet is compared to a resolution function $\mathscr{R}$ for prompt tracks, in order to measure the probability that the track i originates from the primary vertex (Figure 11(a)):

$$\mathscr{P}_i = \int_{-\infty}^{-|d_0^i/\sigma_{d_0}^i|} \mathscr{R}(x)dx \tag{3}$$

The resolution function can be measured in data using the negative side of the signed impact parameter distribution (cf. section 6.5.1), assuming there is no contribution from heavy-flavour particles which is not strictly true.

The individual probability of each of the N tracks associated to the jet are then combined to obtain a jet probability $\mathscr{P}_{jet}$ which discriminates b-jets against light jets (Figure 11(b)):

$$\mathscr{P}_{jet} = \mathscr{P}_0 \sum_{j=0}^{N-1} \frac{(-ln\mathscr{P}_0)^j}{j!} \tag{4}$$

where

$$\mathscr{P}_0 = \prod_{i=1}^{N} \mathscr{P}'_i \quad \text{and} \left\{ \begin{array}{ll} \mathscr{P}'_i = \frac{\mathscr{P}_i}{2} & \text{if } d_0^i > 0 \\ \mathscr{P}'_i = \left(1 - \frac{\mathscr{P}_i}{2}\right) & \text{if } d_0^i < 0 \end{array} \right. \tag{5}$$

Mistagging rate uncertainties (CMS)

Table 1: Average taggability and tagging efficiency in the QCD 80-120 Monte Carlo sample.

flavour		udsg-jets	c-jets	b-jets
taggability (%)		97.2 ± 0.1	98.6 ± 0.1	98.7 ± 0.1
tagging efficiency (%)				
Loose	$TC2 > 2.3$	10.3 ± 0.1	32.8 ± 0.2	71.5 ± 0.2
Medium	$TC2 > 5.3$	0.991 ± 0.008	10.9 ± 0.1	51.0 ± 0.2
Tight	$TC3 > 4.8$	0.108 ± 0.003	2.97 ± 0.06	32.4 ± 0.2

Table 2: Estimated relative systematics (%) on the mistag efficiency for the Track Counting tagger at a jet $p_T = 100$ GeV. Three operating points, corresponding to an average mistag efficiency (in the QCD 80-120 Monte Carlo) of 10%, 1%, 0.1%, respectively, and three luminosity scenarios are considered.

operating point	Loose			Medium			Tight		
Luminosity (pb^{-1})	10	100	1000	10	100	1000	10	100	1000
Systematics (%)									
b fraction	1.4	1.4	0.6	0.8	0.8	0.3	1.2	1.2	0.5
c fraction	0.8	0.8	0.3	0.7	0.7	0.3	1.3	1.3	0.5
g fraction	0.8	0.8	0.4	1.4	1.4	0.7	2.3	2.3	1.2
V^0 fraction	1.4	1.4	0.7	3.6	3.6	1.8	4.6	4.6	2.3
other displaced processes	1.4	1.4	0.7	3.6	3.6	1.8	4.6	4.6	2.3
IP sign flip	0.7	0.3	0.2	4.5	1.9	1.4	24.0	10.2	7.6
statistics MC	0.1	0.1	0.1	0.4	0.4	0.4	1.2	1.2	1.2
statistics data	0.4	0.1	—	1.6	0.5	0.2	5.5	1.7	0.6
sampling	2.0	2.0	2.0	5.0	5.0	5.0	13.0	13.0	13.0
Total syst.	3.4	3.4	2.4	8.8	7.6	5.9	28.7	18.1	15.5

Efficiency determination uncertainties (CMS)

	TrackCounting	TrackProbability
	Loose	Loose
pTrel (n)	0.68 ± 0.01 (stat) ± 0.10 (Ts)	0.71 ± 0.01 (stat) ± 0.11 (Ts)
pTrel (p)	0.71 ± 0.01 (stat) ± 0.11 (Ts)	0.74 ± 0.01 (stat) ± 0.11 (Ts)
Counting	0.69 ± 0.01 (stat) ± 0.10 (Ts) ± 0.02 (Ms)	0.71 ± 0.01 (stat) ± 0.11 (Ts) ± 0.03 (Ms)
MC truth	0.69 ± 0.01	0.73 ± 0.01
	Medium	Medium
pTrel (n)	0.51 ± 0.01 (stat) ± 0.08 (Ts)	0.47 ± 0.01 (stat) ± 0.07 (Ts)
pTrel (p)	0.53 ± 0.01 (stat) ± 0.08 (Ts)	0.50 ± 0.01 (stat) ± 0.07 (Ts)
Counting	0.52 ± 0.01 (stat) ± 0.08 (Ts) ± 0.02 (Ms)	0.46 ± 0.01 (stat) ± 0.02 (Ts) ± 0.02 (Ms)
MC truth	0.50 ± 0.01	0.48 ± 0.01
	Tight	Tight
pTrel (n)	0.30 ± 0.01 (stat) ± 0.05 (Ts)	0.29 ± 0.01 (stat) ± 0.04 (Ts)
pTrel (p)	0.26 ± 0.01 (stat) ± 0.04 (Ts)	0.28 ± 0.01 (stat) ± 0.04 (Ts)
Counting	0.32 ± 0.01 (stat) ± 0.05 (Ts) ± 0.01 (Ms)	0.28 ± 0.01 (stat) ± 0.04 (Ts) ± 0.01 (Ms)
MC truth	0.29 ± 0.01	0.30 ± 0.01

Table 1: Summary of semileptonic b-tagging efficiencies measured with the pTrel and Counting methods compared with those obtained from MC truth. The systematic uncertainty on the p_{Trel} templates is shown in the table as (Ts). The uncertainty on the mistag rate in the Counting method is shown as (Ms). The uncertainty on the MC truth efficiency is only statistical.