B-tagging at the Tevatron

Flera Rizatdinova (KSU)

- *Motivation*
- *Overview of b-tagging methods*
- *Measurement of b-tagging efficiency and mistagging rate*
- *b-tagging performance at the Tevatron*
- *Example of using b-tagging in top cross section measurement*
- *b-tagging at the LHC*
- *Conclusions*

Motivation

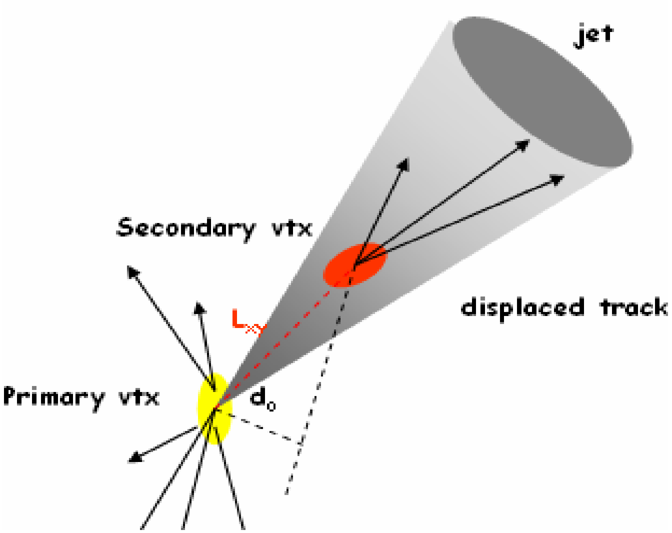
- B-tagging (b-jet identification) plays a crucial role for the high energy physics:
  - Top physics – all aspects (Br(t→Wb)~100%)
  - Higgs physics – search for associated Higgs production
    - (W/Z)H, H→*bb*
    - H*tt,* H→*bb*
  - Searches for new physics
    - Stop, sbottom searches;
    - Technicolor searches
- Most of these applications require high b-tagging efficiency (probability to tag a b-jet) ~50—60% while keeping mis-tagging rate (probability to tag light jets originated from u/d/s-quarks and gluons) below 1%

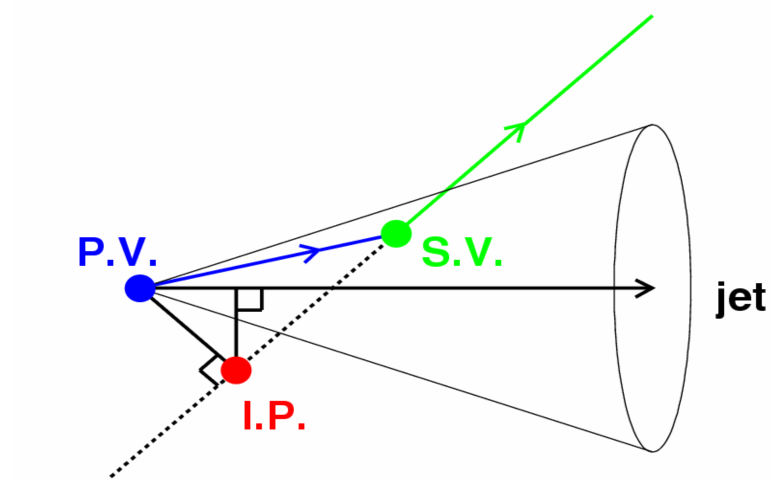
B-tagging algorithms (1)

- Soft lepton tagging (SLT)
  - Based on the fact that ~20% of *b*-hadrons decay semileptonically (either *e* or μ in the final state)
  - Usually do not use electrons – big challenge;
  - Once having preselected jets with muons, SLT has high efficiency, but high mistag rate as well
  - Purity of this tagger can be improved using kinematic variables of muon, for example, using p_{Trel} – p_T of muon with respect to jet axis (muons from b-decays tend to have larger p_{Trel})

B-tagging algorithms (2)

- Lifetime tagging
  - Make use of relatively big lifetime of B-hadrons
  - Two different techniques
    - Explicit reconstruction of secondary vertices in jets
    - Looking at tracks in jets which are significantly displaced from primary vertex (counting method or probabilistic approach)





Lifetime b-tagging methods

Secondary Vertex Tag

- Look for displaced vertices
- Jet is tagged as a *b* jet if signed decay length significance > some value (depends on "acceptable" mistag rate)

Counting Signed Impact Parameter tag (CSIP)

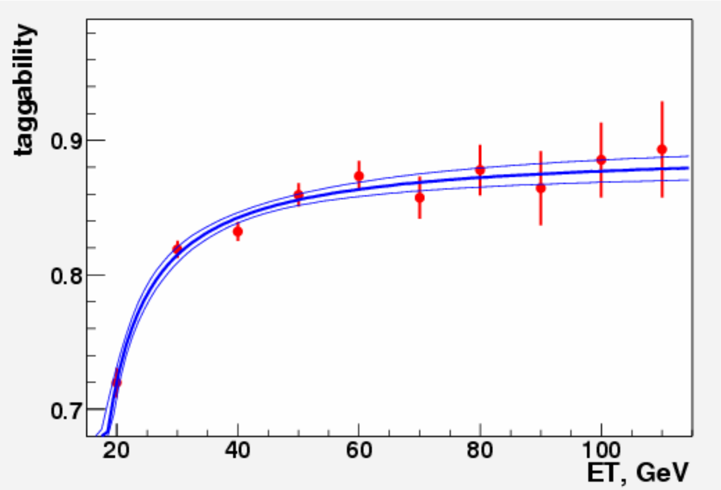
- $S = IP/\sigma(IP)$
- Jet is positively tagged if it has
  - at least two tracks with S>3 or
  - at least three tracks with S>2

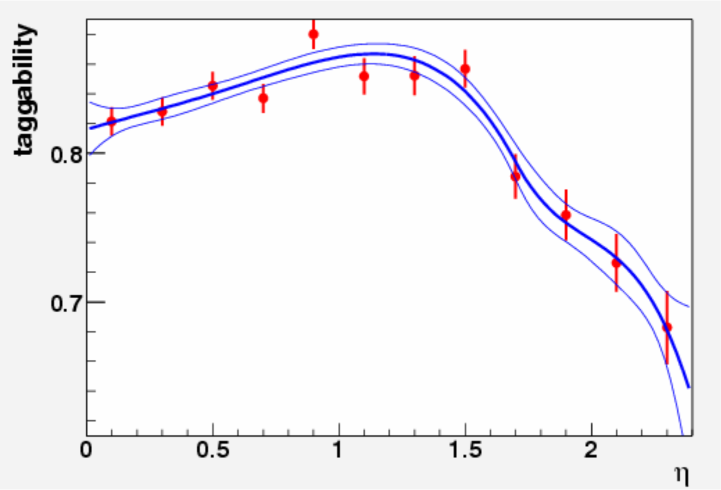
Jet Lifetime Probability

- For each track in the jet calculate a probability to come from primary vertex based on the IP significance;
- Combine probabilities for individual tracks into jet probability;
- Jet is tagged if its probability to be a light jet is less than some value (depends on "acceptable" mistag rate)

Taggability

- The idea is to largely decouple the tagging efficiency from issues related to tracking inefficiencies and/or calorimeter noise;
- Is used by both experiments;
- Taggability is the probability for a jet to pass some quality criteria (D∅):
  - At least two tracks in track-jet associated with calorimeter jet
  - Certain criteria on p_T of tracks and on the number of hits per track
- Found that it depends on event sample and is not fully modeled by Monte Carlo





How to measure b-tagging efficiency on data?

- Ideal case (or Monte Carlo) : tag a sample of b-jets, so efficiency is the ratio of tagged jets to the total number of taggable jets
- Data: mixture of light (dominant), c- and b-jets
  - Both numerator and denominator are affected by presence of non-b jets
- Need separation between b and non-b jets
- Requirement of soft lepton increases probability for a jet to be heavy flavor

- CDF: e-in-jets
  - $p_T(e)>8$ GeV
  - Di-jet events
- D∅: μ-in-jets
  - $p_T(\mu)>8$ GeV
  - Di-jet events

Double tag method (CDF)

- Enrich the di-jet sample with heavy flavor by tagging away jet;
- Subtract the number of mistags (tagged light jets) assuming it equal to the number of negative tags;
- The result is corrected for the fraction of non-QQbar events
  - Tagged away jet is a mistagged light jet or it comes from gluon splitting
  - Electron is fake or part of a conversion
- The correction factor involves
  - Efficiency to reconstruct conversions – measured on data
  - b/c ratio – measured on data, from D^0 meson invariant mass fit or looking at μ from cascade c-decays in e-jet.
- *b*-tagging efficiency:

$$\varepsilon = \frac{(N_{a+}^{e+} - N_{a+}^{e-}) - (N_{a-}^{e+} - N_{a-}^{e-})}{N_{a+} - N_{a-}} \cdot F_B^a$$

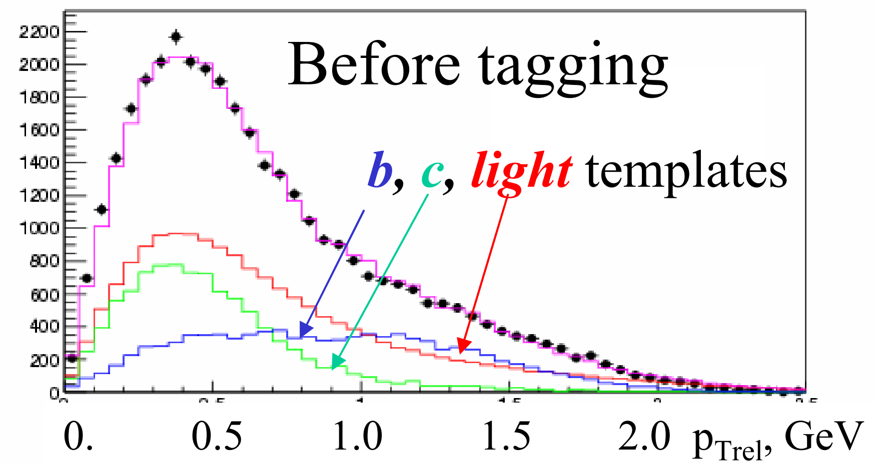
P_{Trel} method (D∅)

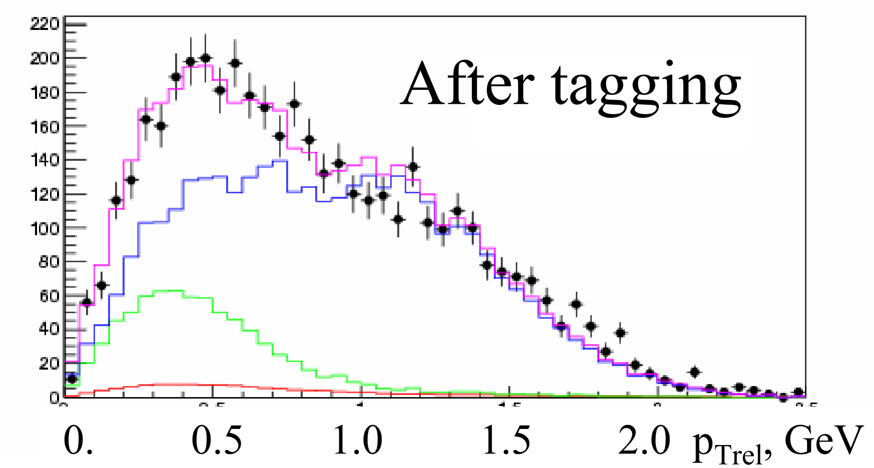
- Look at muonic jets
- p_{Trel} is different for μ originated from light, c and b-decays – use this fact to measure fraction of b-jets before tagging ($F_{b\to\mu}$) and after tagging ($F^{tag}_{b\to\mu}$)

Efficiency is calculated as

$$\varepsilon_{btag} = \frac{N^{tag}_{\mu} \cdot F^{tag}_{b\to\mu}}{N_{\mu} \cdot F_{b\to\mu}}$$

- Light template – obtained from data;
- *b, c* templates obtained on MC





System 8 (basic method at D∅)

- Have two samples with different heavy flavor fractions
  - μ-in-jet sample (n-sample)
  - The same sample with tagged away-jet (p-sample)
- Tag muonic jets with two independent tagging algorithms
  - Lifetime (LT = JLIP, CSIP or SVT)
  - SLT requiring $p_{Trel} > 0.7$ GeV
- Write a system of eight equations to solve for ε_{btag}^{LT}

$$n = n_b + n_l$$

$$p = p_b + p_l$$

$$n^{LT} = n_b \varepsilon_{btag}^{LT} + n_l \varepsilon_{non-b}^{LT}$$

$$p^{LT} = p_b \varepsilon_{btag}^{LT} + p_l \varepsilon_{non-b}^{LT}$$

N of jets tagged by lifetime methods

$$n^{SLT} = n_b \varepsilon_{btag}^{SLT} + n_l \varepsilon_{non-b}^{SLT}$$

$$p^{SLT} = p_b \varepsilon_{btag}^{SLT} + p_l \varepsilon_{non-b}^{SLT}$$

N of jets tagged by soft lepton tag

$$n^{dt} = n_b \varepsilon_{btag}^{LT} \varepsilon_{btag}^{SLT} + n_l \varepsilon_{non-btag}^{LT} \varepsilon_{non-b}^{SLT}$$

$$p^{dt} = p_b \varepsilon_{btag}^{LT} \varepsilon_{btag}^{SLT} + p_l \varepsilon_{non-btag}^{LT} \varepsilon_{non-b}^{SLT}$$

N of jets tagged by both soft lepton tag and lifetime method (double tags)

Assume: *decorrelation between taggers; cross checked it on MC, decorrelation factor 0.98*

Main assumptions for ε_{btag}

- Double tag, System8 methods:
  - Probability to tag a jet with soft lepton does not depend on probability to tag away jet – not a big effect;
  - There are jets originated from $g \rightarrow bb$, with both *b*-quarks in one jet – these jets presumably have higher tagging efficiency;
  - Assume that gluon splitting contribution in MC is correct, both in shape and in size
- Double tag specific:
  - b/c ratio is constant as function of jet p_T

Main assumptions for ε_{btag} (cont'd)

- System8 specific:
  - Have *b* and non-*b* efficiencies: p_{Trel} (used to enhance HF fraction) templates do not distinguish between *c* and light jets;
  - Cross-check with other *b*-tagging measurements and MC– got the same result within errors;
- p_{Trel} fit method (templates obtained from MC):
  - correct model of *b*-decays – checked with different generators, moderate effect;
  - correct fragmentation model – almost no effect

Experimental limitation now – low statistics at high jet p_T, so have to rely on MC shape of b-tagging efficiency

b,*c*-tagging efficiencies in data and MC

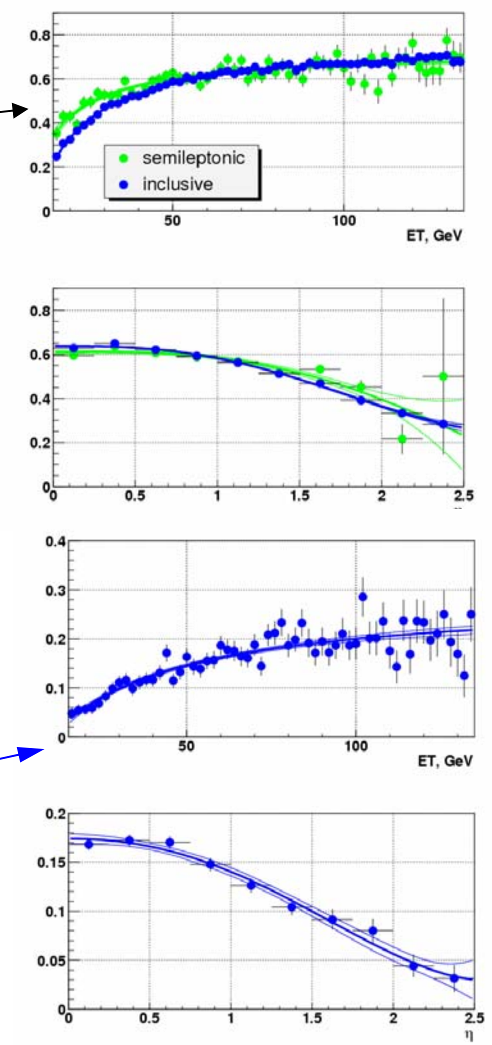
Both CDF and D∅ measure semileptonic *b*-tagging efficiency on data;

Semileptonic *b*-tagging efficiency in MC is higher than measured in data – need to introduce a scale factor between data and MC (CSIP at D0: 0.75; SecVtx at CDF: 0.82):

$$SF_{b\to\mu} = \varepsilon^{data}_{b\to\mu} / \varepsilon^{MC}_{b\to\mu}$$

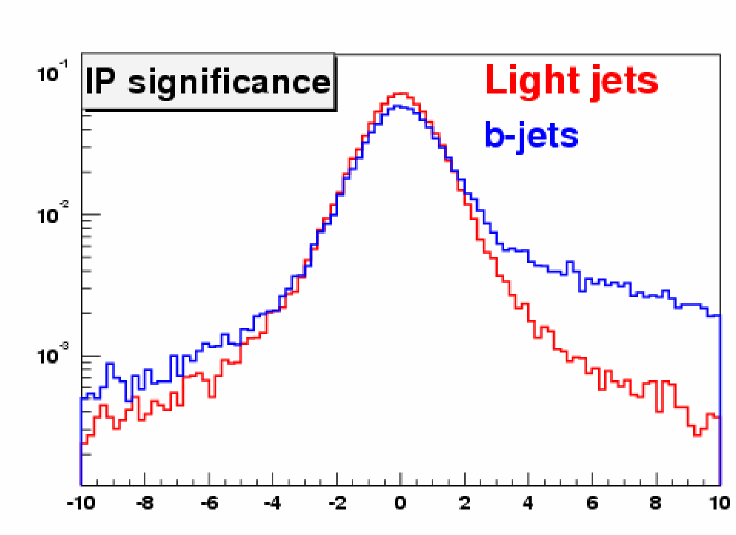
Need inclusive *b*-tagging efficiency → assume the same data/MC ratio as for semileptonic jets

Do not measure *c*-tagging efficiency on data – instead, use MC *c*-tagging efficiency corrected by ratio of *b*-tagging efficiencies in data to MC.



Measurement of mistagging rate in data

- Measure negative inclusive tagging rate on data
  - Use negative side of DCA significance distribution (impact parameter based methods) or negative decay length (secondary vertex algorithms)
- Need to convert negative tagging rate into light jet tagging rate

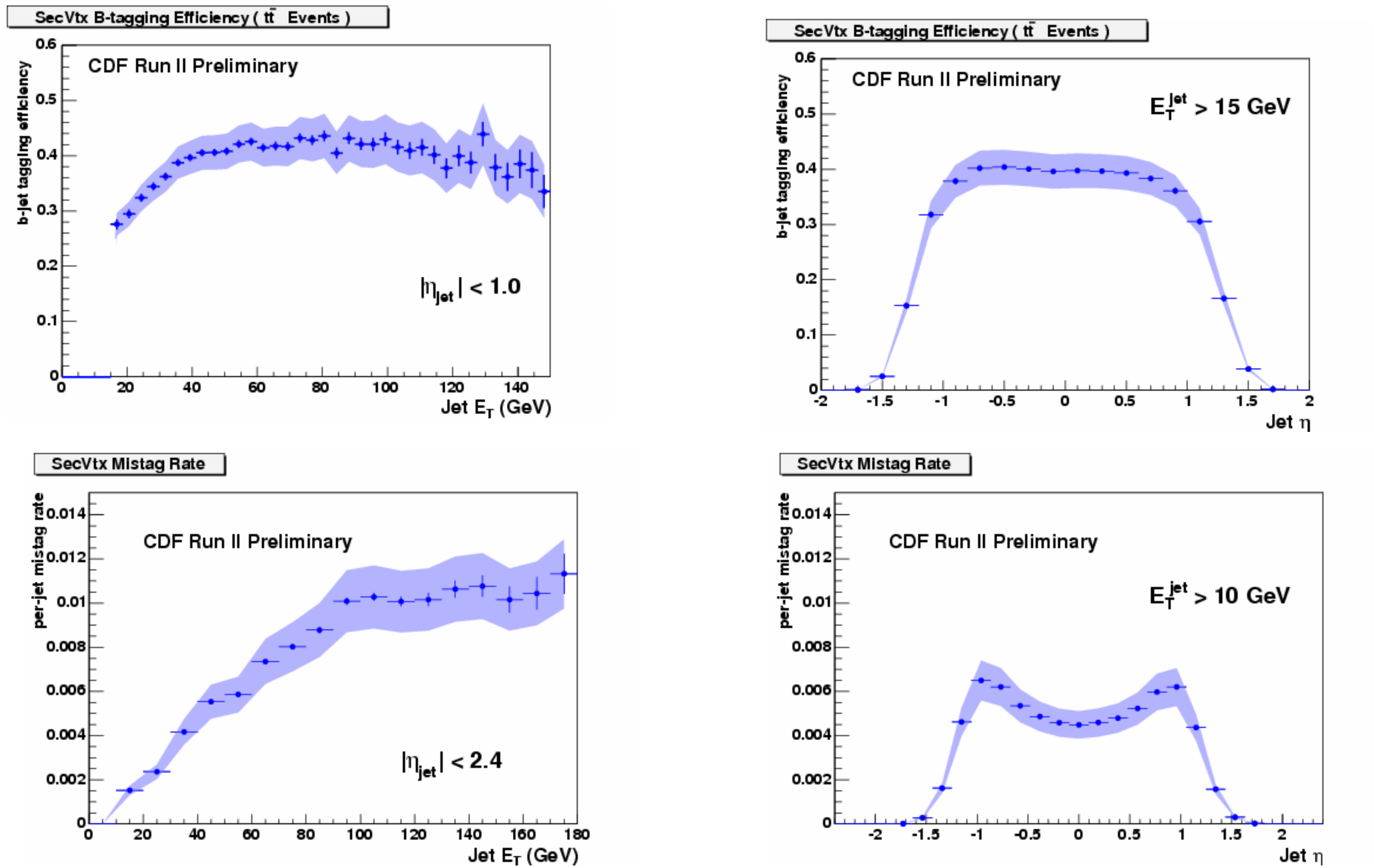


Measurement of mistagging rate in data

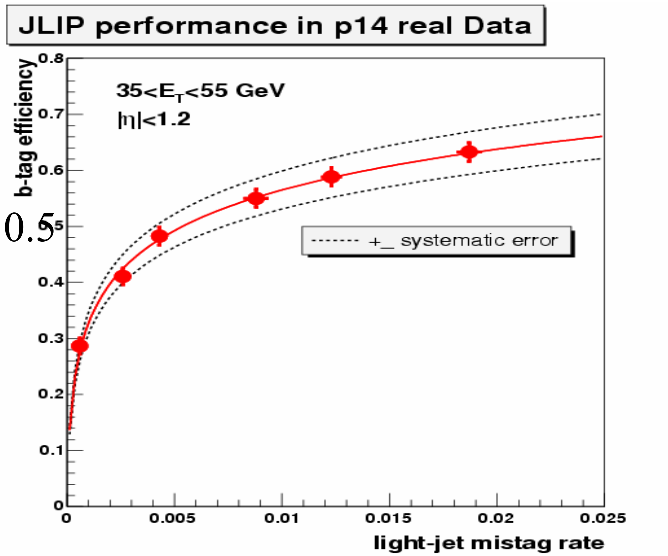
- Difference between light positive tag rate (mistags) and measured negative tag rate:
  - Fake tracks, long-lived particles not cleaned up by V0 filters, interactions with material give rise to positive tag rate – need scale factor SF_{ll};
    - Both experiments rely on MC for estimation of SF_{ll};
    - CDF has measured correction factor due to material interactions directly on data
  - Presence of heavy flavor in the negative tail artificially increases negative tag rate – introduce another scale factor SF_{hf}
- Finally, mistagging rate is:

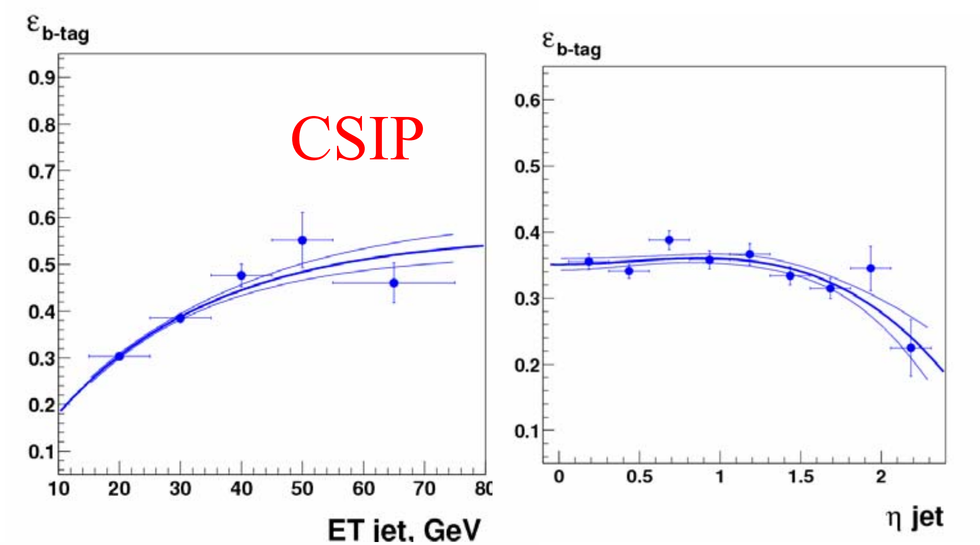
$$\varepsilon_{light}(x, y, \ldots) = \varepsilon^{-}(x, y, \ldots) \cdot SF_{hf} \cdot SF_{ll}$$

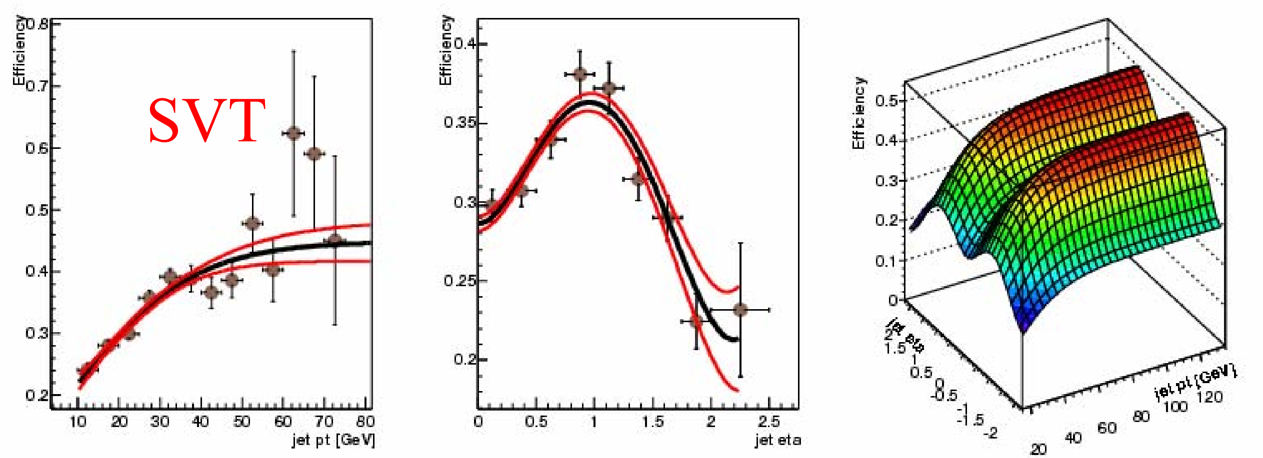
b-tagging performance on data (CDF)



Results: *b*-tagging efficiency in data (D∅)





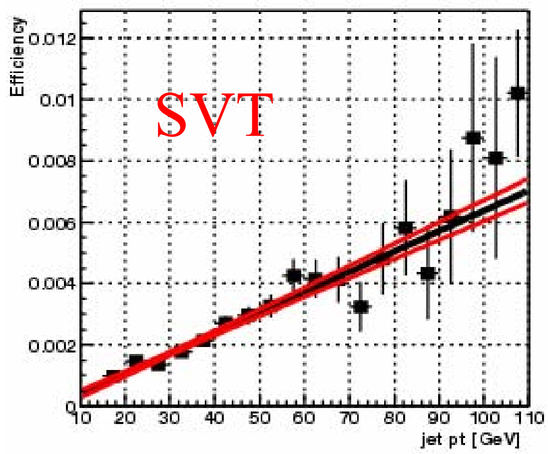


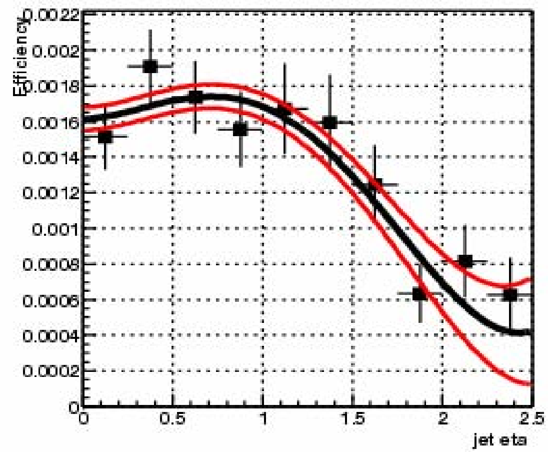
Probability to tag a *tt* event P(n_{tag} ≥ 1):

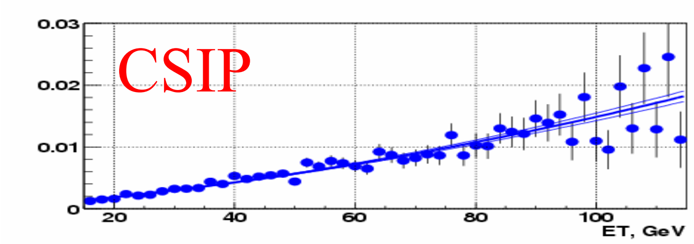
CSIP: ~61%

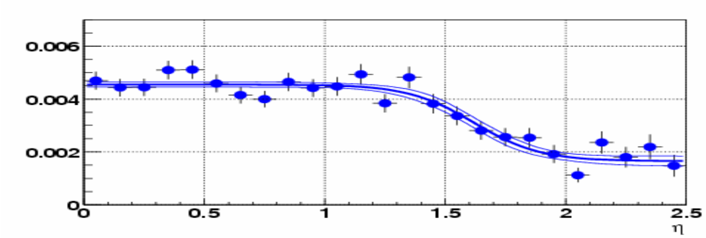
SVT : ~58%

Mistagging rate in data (D∅)









W+4 light jets events

P(n_{tag}≥1):

SVT: ~1.1%

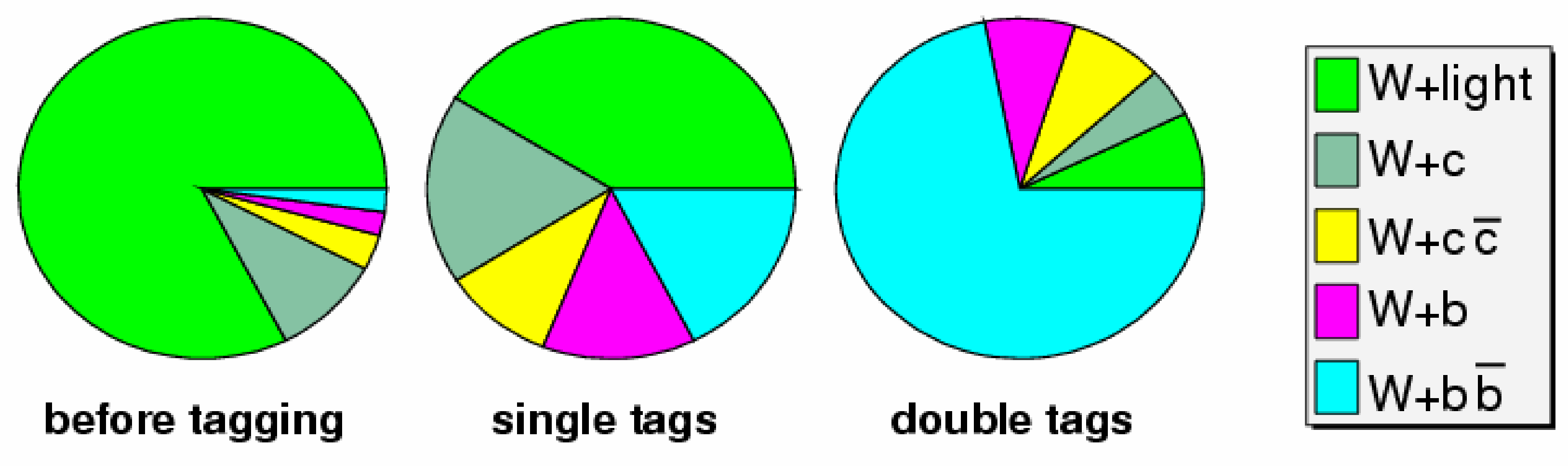
CSIP: ~2.6%

Role of *b*-tagging in ttbar studies

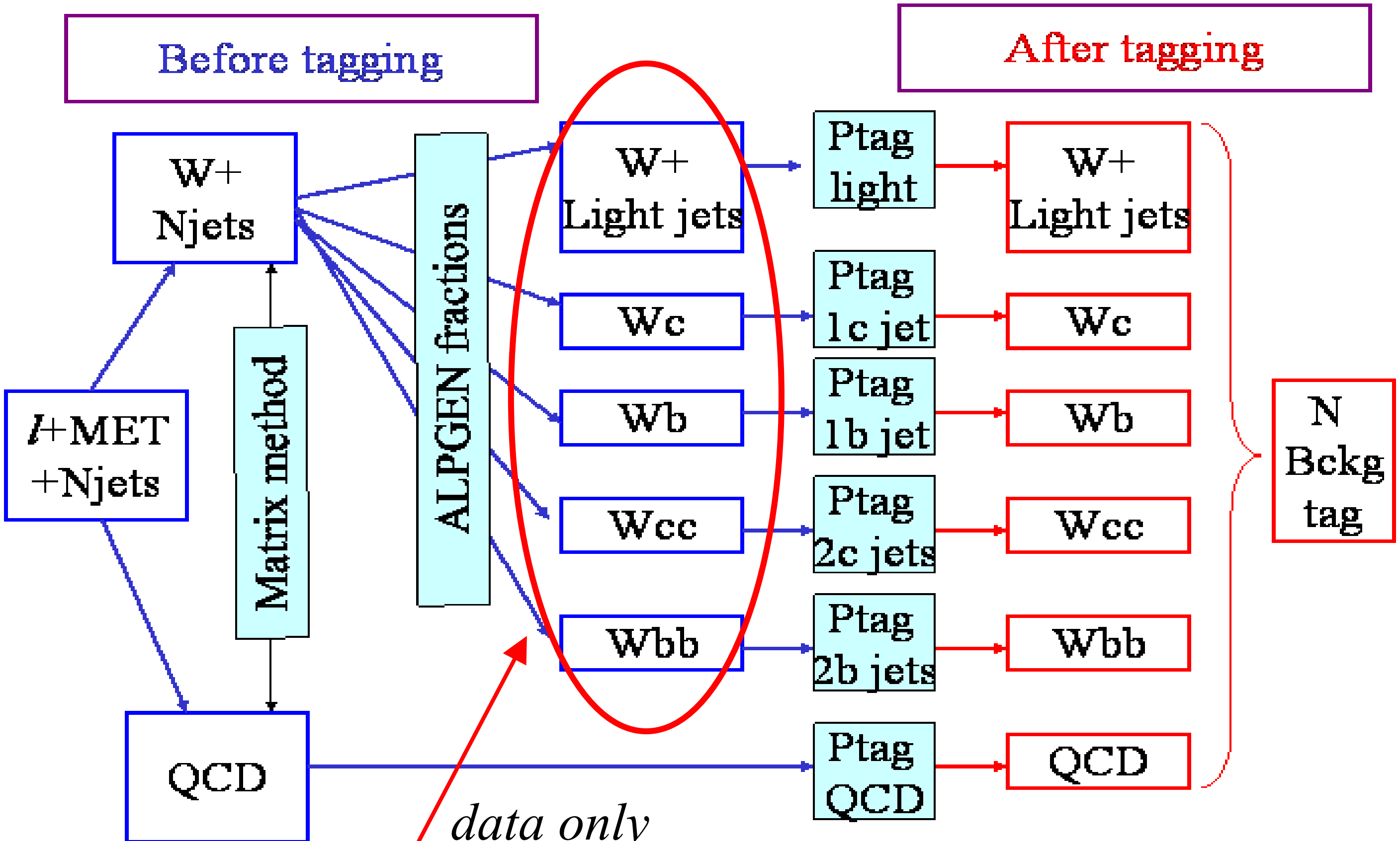
Why do we want to use b-tagging in top analyses?

After suppressing QCD background by requirement of high MET and “real” isolated lepton, W+jets production is the main BG.

W+jets composition in the 4th jet multiplicity bin:



b-tagging in top cross section measurement



Estimate tt production cross-section from the excess observed in the number of tagged events w.r.t. BG expectation in 3 and 4jet multiplicity bins.

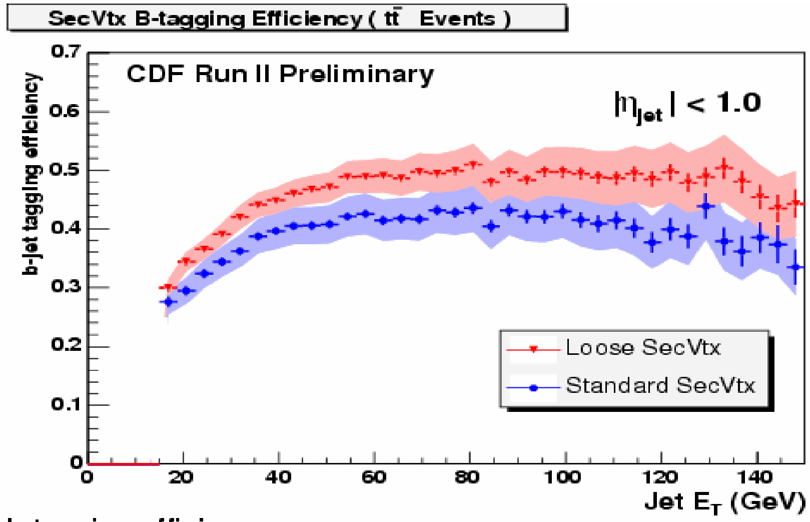
One of the most important inputs from theory

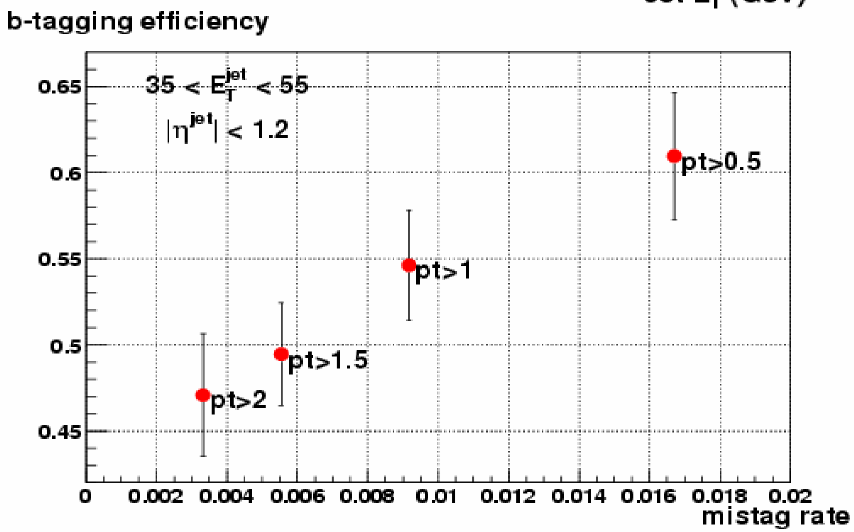
Estimation of W+jets background

- Use *W+jets* sample generated with ALPGEN (LO ME) interfaced to PYTHIA;
- Apply matching procedure (D∅ - ad hoc) to eliminate double counting and reduce sensitivity to parton generation cuts;
  - Want to have well defined procedure which works for all types of quarks
  - Need predictions for fractions from NLO calculations, for instance, we can correct LO predictions by NLO scale factor using W+2 jets calculations
- Rely on ratios of the cross sections – apply 50% uncertainty;
  - Is it too conservative?
- Use MC *b,c*-tagging efficiencies corrected for the difference between data and MC using scale factors;
- Rely on MC kinematics of *W+n jets* (and ttbar) events when calculate average event tagging probability
  - cross checks (after tagging) of kinematical distributions made by CDF and D∅ show that kinematics is described reasonably (at the level of available statistics)

b-tagging R&D at the Tevatron

- Increase b-tagging efficiency by the price of increasing mis-tagging rate (CDF&D∅)
  - reasonable for cases where the dominant background is expected to be heavy flavor production
- Increase acceptance of b-tagging algorithms (CDF)
- Suppress c-tagging efficiency (done by CDF, work is going on in D∅)
  - Important for further background rejection in top quark studies, Higgs boson and some signatures of new physics searches.





b-tagging at the LHC

- Different vertex detector – pixel tracker, much better single hit position resolution
  - But larger inner radius R_{inner} compared to CDF, D∅ detectors – b-tagging efficiency is not expected to be essentially higher than at the Tevatron;
- Best sample to measure b-tagging efficiency – ttbar events! (in addition to the standard lepton-in-jet sample)
  - For example, dilepton channel requiring one b-tag – huge statistics, high purity of b-jet sample
- Will have a possibility to measure c-tagging efficiency directly on data:
  - ttbar events, lepton+jets sample, look at the hadronic decays of W boson (no b-jets)
  - good jet energy resolution to reconstruct hadronic W in top events.

Conclusions

- CDF and D∅ have very similar b-tagging performance on data;
- Both experiments have well developed methods to measure b-tagging efficiency and mistagging rates on data;
- Further studies on *b*-tagging are going in the same direction;
- Hope our experience will be useful for LHC.